SIGTYP 2025

The 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP

Proceedings of the Workshop

August 1, 2025

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-281-7

Introduction

We are pleased to present the proceedings of SIGTYP 2025, the seventh edition of the Workshop on Research in Computational Linguistic Typology and Multilingual Natural Language Processing. This year, the workshop is held as a joint event with FieldMatters and is co-located with the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), taking place in Vienna, Austria.
Building on the success of previous editions from 2019 through 2024, SIGTYP continues to serve as a platform for fostering dialogue between the fields of linguistic typology and multilingual NLP. Our core mission remains the same: to raise awareness of typological diversity and to explore how insights from linguistic typology can inform, enrich, and challenge computational methods in cross-lingual and multilingual settings. We are particularly committed to the development of truly inclusive NLP methods that serve a broad and typologically diverse range of languages.
SIGTYP 2025 invites contributions at the intersection of typology and NLP, with key areas of focus including:

- The integration of typological features in multilingual learning and language transfer;
- The development of unified linguistic taxonomies and cross-lingual resources;
- Automatic inference of typological features using machine learning;
- Enhancing interpretability of multilingual models through typological knowledge;
- Collaborative approaches to improving typological databases;
- Addressing the challenges of cross-lingual annotation and defining linguistic universals;
- Language-specific studies aimed at supporting or revising typological claims.

This year's program includes 2 keynote talks, 15 archival papers and 2 extended abstracts. We are honored to host Robert Forkel and Lisa Bylinina as invited speakers, whose work exemplifies the interdisciplinary spirit of the workshop. We extend our sincere thanks to all authors for their high-quality submissions, to the program committee for their diligent and insightful reviews, and to all participants who contribute to the vibrancy and impact of SIGTYP. For more information, including proceedings and shared task resources, please visit the workshop website: website: https://sigtyp.github.io/ws2025-sigtyp.html

Organizing Committee

Organizing Committee

Michael Hahn, Saarland University
Priya Rani, University of Galway
Ritesh Kumar, Dr. Bhimrao Ambedkar University
Andreas Shcherbakov, The University of Melbourne
Alexey Sorokin, Yandex and Lomonosov Moscow State University
Oleg Serikov, King Abdullah University of Science and Technology
Ryan Cotterell, Swiss Federal Institute of Technology
Ekaterina Vylomova, The University of Melbourne

Program Committee

Program Chairs

Michael Hahn, Saarland University
Priya Rani, University of Galway
Oleg Serikov, King Abdullah University of Science and Technology
Andreas Shcherbakov, University of Melbourne
Ritesh Kumar, Dr. Bhimrao Ambedkar University
Alexey Sorokin, Yandex and Lomonosov Moscow State University
Ekaterina Vylomova, The University of Melbourne
Ryan Cotterell, Swiss Federal Institute of Technology

Reviewers

Emily Ahn

Barend Beekhuizen, Claire Bowern

Giuseppe G. A. Celano, Ryan Cotterell, Jannic Alexander Cutura

Rena Wei Gao

Michael Hahn, Borja Herce

Elisabetta Jezek

Ritesh Kumar, Kemal Kurniawan

M. Dolores Jiménez López

Aso Mahmudi, Raphael Merx

Gaurav Negi

Devishree Pillai, Edoardo Ponti

Priya Rani

Oleg Serikov, Andreas Shcherbakov, Alexey Sorokin, Richard Sproat

Ekaterina Vylomova

Jinrui Yang

Keynote Talk

Connecting the dots - growing an eco-system for cross-linguistic data

Robert Forkel
Max Planck Institute for Evolutionary Anthropology
2025-08-01 – Room: **TBD**

Abstract: One of the key contributions typology can make to multilingual NLP is a fuller picture of the diversity of the world's languages. This diversity is also reflected in widely varying documentation across languages. Thus, informing computational approaches to language processing by this diversity requires operationalizing a variety of data types describing very different languages. Getting a computational grasp on cross-linguistic information has been the main motivation behind CLDF - the Cross-Linguistic Data Formats. This talk will explore the eco-system of cross-linguistic data that is now opened up via CLDF.

Bio: Robert Forkel leads the research data management group and serves as a scientific programmer in the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology in Germany. His current work centers on developing software solutions to collect, curate, and publish large-scale databases for linguistic and cultural research. He is also interested in the role of data in scientific research, with a particular focus on reproducibility. In addition, he contributes to open-source software packages such as LingPy.

Keynote Talk

(L)LMs and language theory

Lisa Bylinina
Utrecht University
2025-08-01 – Room: **TBD**

Abstract: One of the central questions in linguistic typology is: What constrains the space of natural languages? In a somewhat narrower formulation: How do different grammatical properties of a language relate to each other, and why are some combinations of features that would, in principle, be possible, in fact not attested? I would like to put these questions in the context of recent language models. Can (L)LMs help us understand interconnections within linguistic grammatical systems? I will argue for a moderately optimistic view and suggest some ways to make progress in this direction, with a focus on the linguistic generalisations (L)LMs make under different training conditions. My goal is to encourage discussion about the usefulness of (L)LMs for theoretical and typological linguistic research.

Bio: Lisa Bylinina is an Assistant Professor of Computational Linguistics (UD1) at Utrecht University, where she is part of the Language and Communication group within the Institute for Language Sciences. She is also an active member of the NLP@U special interest group. Her research interests lie at the intersection of theoretical linguistics and natural language processing. Before joining Utrecht University in September 2024, she held the position of Assistant Professor at the University of Groningen, in the Computational Linguistics Group at the Center for Language and Cognition (CLCG). At Utrecht, she teaches in the Applied Data Science master's program and the bachelor's program in Communication and Information Science. She is open to supervising (research) master's theses in data science, artificial intelligence, and theoretical linguistics, particularly in semantics.

Table of Contents

InstructionCP: A Simple yet Effective Approach for Transferring Large Language Models to Target Languages

Kuang-Ming Chen[1,2]* **Jenq-Neng Hwang**[1] **Hung-yi Lee**[3]
[1]University of Washington, Seattle, WA, USA
[2]ASUS Open Cloud Infrastructure Software Center, Taipei, Taiwan
[3]National Taiwan University, Taipei, Taiwan
kmchen@uw.edu hwang@uw.edu hungyilee@ntu.edu.tw

Abstract

The rapid development of large language models (LLMs) in recent years has largely focused on English, resulting in models that respond exclusively in English. To adapt these models to other languages, continual pre-training (CP) is often employed, followed by supervised fine-tuning (SFT) to maintain conversational abilities. However, CP and SFT can reduce a model's ability to filter harmful content. We propose Instruction Continual Pre-training (InsCP), which integrates instruction tags—also known as chat templates—into the CP process to prevent loss of conversational proficiency while acquiring new languages. Empirical evaluations on language alignment, reliability, and knowledge benchmarks confirm the efficacy of InsCP. Notably, this approach requires only 0.1 billion tokens of high-quality instruction-following data, thereby reducing resource consumption.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across numerous natural language processing (NLP) tasks(Brown et al., 2020). However, the majority of LLMs are pre-trained on English corpora(AI@Meta, 2024; Team et al., 2024; OpenAI, 2023), thus restricting their utility to English language contexts.

While some endeavors opt to train their LLMs from scratch using non-English data, as exemplified by YI-34B(AI et al., 2024), we recognize the significant time and computing resources required for such an approach. Drawing inspiration from Ouyang et al. (2022), many research groups have shifted their focus towards continual pre-training (CP)(Gupta et al., 2023; Ke et al., 2022) on target languages to enhance knowledge acquisition and model fluency. Subsequently, supervised fine-tuning (SFT) is conducted on instruction-formatted data to ensure that models possess the capability to respond to questions in a format consistent with English-based pre-trained LLMs, such as BLOOM(Workshop et al., 2023), LLaMA2(Touvron et al., 2023), and Mistral-7B(Jiang et al., 2023).

Yet, as highlighted in Qi et al. (2023), challenges persist in maintaining RLHF capabilities when fine-tuning GPT-3.5 turbo(OpenAI, 2023) on non-English data. Our experiments validate similar observations with other LLMs like LLaMA2.

This work proposes a novel fine-tuning approach called Instruction Continual Pre-training (InsCP) for LLMs to adapt to non-English languages. We hypothesize that providing a chat template during CP prevents the model from forgetting its conversational abilities, as it mirrors its original training conditions. InsCP is essentially the same as typical CP, except that we augment each piece of data with a chat template containing special instruction tokens, such as $< |begin\_of\_text| >$ in LLaMA3(AI@Meta, 2024). This simple augmentation enables the model to effectively retain its original RLHF capabilities, such as defending against offensive input while learning a new language through CP.

We evaluate the effectiveness of InsCP on LLMs, primarily focusing on the LLaMA3-instruct model, across three key aspects: language alignment, reliability, and knowledge benchmarks.

The results demonstrate that the model, after undergoing InsCP on LLaMA3-instruct, effectively performs in Traditional Chinese when prompted with Traditional Chinese input, surpassing the performance of LLaMA3-instruct. Moreover, the model after InsCP does not suffer a serious performance dropped in knowledge, safety and RLHF ability.

*This work was conducted while the first author was an intern at ASUS Open Cloud Infrastructure Software Center.

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 1–6
August 1, 2025 ©2025 Association for Computational Linguistics

2 Related Work

2.1 LLMs adapt in other languages

Fine-tuning is a widely-used technique for adapting models, particularly in the domain of large language models (LLMs), to specific domains. Many downstream tasks have been successfully addressed through fine-tuning (Howard and Ruder, 2018; Devlin et al., 2019; Radford et al., 2018). While most downstream tasks can be accomplished through supervised fine-tuning, adapting an English-based LLM to other languages, such as in the work of Fujii et al. (2024); Zhao et al. (2024); Cui et al. (2023); Lin and Chen (2023); YuLan-Team (2023) for non-English languages, typically begins with continual pre-training (CP). This initial step is crucial for ensuring that the models possess the necessary language proficiency and knowledge. Since acquiring proficiency in a specific language requires a large amount of data, CP is advantageous as it does not require labeled data, enabling the use of vast amounts of available language data. Subsequently, instruction fine-tuning allows the model to engage in conversational interactions using specific templates.

2.2 Problems of Continual Pre-training

Continual pre-training (CP) is often employed to adapt those English-based models to other languages. However, Li and Lee (2024) points out that CP can lead to catastrophic forgetting, particularly diminishing the model's conversational abilities. To address this issue, Huang et al. (2024) proposed a method called "chat vector," which enhances chat capabilities through model weight arithmetic, achieving good performance across various benchmarks. Despite these advancements, many researchers continue to tackle the challenges posed by CP. In this work, we present a straightforward approach to mitigate these issues.

3 Methodology

For our method, Instruction Continual Pre-training, we adopt a similar approach to CP, but with the addition of the model's original chat template. The template is shown in Appendix A.1 The **inputs** in the template represent the prompts provided by the user. In our context, where the objective is to train LLMs in the target language through next token prediction tasks while retaining their chat ability, we place the CP data in the **model_response**. This arrangement ensures that LLMs generate tokens based on the target language. The InsCP template is shown in A.1.

4 Experimental Setup

4.1 Pre-training Dataset

We utilize a high-quality dataset comprising paired instruction-following data for LLaMA3-instruct 8B(AI@Meta, 2024) during the InsCP procedure. The InsCP procedure means the traditional CP method with instruction-following data. The dataset consists of Traditional Chinese text and has a total size of 0.1 billion tokens. Throughout the InsCP process, we segregate the questions and answers into two separate data points. Further details regarding the training process are provided in the Appendix A.3.

Moreover, to demonstrate the generalizability of our method to other languages, we extend our approach to Japanese. We utilize a 70M tokens dataset, which is also instruction-following data same as the Traditional Chinese dataset structure, to perform InsCP on LLaMA3-instruct 8B.

From our experiments, we discovered the critical importance of selecting appropriate data for InsCP. We aimed to determine the most suitable type of data for InsCP. Based on our findings, we selected instruction-following data with low perplexity because low perplexity are likely to closely resemble the original output of LLMs, thereby minimizing any adverse effects on the models' original abilities.

4.2 Evaluation

4.2.1 Language Alignment

To evaluate language alignment, we employ the FastText language identification model (Joulin et al., 2016a,b). This model is used to determine the language of 2000 aligned sentences extracted from the English and Traditional Chinese subset of the NeuLab-TedTalks language within the tokens generated by our model. The FastText model classifies text into two categories: Chinese and English. The results include the percentage of sentences identified as Chinese, English, and others from the set of 2000 input prompts.

4.2.2 Reliability

We assess the reliability of the model's output using several common benchmarks, including TruthfulQA(Lin et al., 2022), ToxiGen(Hartvigsen et al., 2022), and BOLD(Dhamala et al., 2021), utilizing lm-evaluation-harness(Gao et al., 2021).

4.2.3 Knowledge Benchmarks

We utilize several benchmarks to evaluate our model's knowledge: **C-eval-tw**: A translation of C-eval(Huang et al., 2023), used to evaluate our model. Compute metrics by averaging accuracy across individual tasks. The accuracy computation involves selecting the option with the highest probabilities. **TTQA**(Hsu et al., 2023): Focuses on Taiwanese commonsense and knowledge by using 64 expert-selected paragraphs from Wikipedia. We extract the model's output and calculate accuracy based on multiple-choice questions. **TMMLU Plus**(Tam et al., 2024): Used for traditional Chinese multitask benchmarking. We calculate accuracy for each task directly. **ARC**(Clark et al., 2018) and **Hellaswag**(Zellers et al., 2019): Ensure that our model's English-related knowledge does not degrade. We utilize length-normalized accuracy. **MMLU**(Hendrycks et al., 2020): Suitable for multitask evaluation. We calculate accuracy for each task directly.

4.2.4 MT-Bench

MT-Bench(Zheng et al., 2023) incorporates multi-conversation scenarios, allowing us to assess the model's ability to handle multiple interactions simultaneously. This enables us to demonstrate that InsCP does not compromise the RLHF ability of the model. In MT-Bench, the GPT-4 score serves as our evaluation metric, and we include a prompt about judging language alignment in GPT-4 evaluation to test the model's language ability.

4.3 Baselines

We select LLaMA-3-instruct as our baseline model. To evaluate the performance of Instruction Continual Pre-training (InsCP), we conduct InsCP using our baseline model. Importantly, it's worth noting that both InsCP and the original continual pre-training (orgCP) utilize the same continual pre-training (CP) data. Furthermore, to compare with the original continual pre-training process, we also fine-tune a model using original continual pre-training.

Model	**EN Prompt**		**ZH Prompt**	
	EN% ↑	**ZH% ↓**	**EN% ↓**	**ZH% ↑**
LLaMA3-instruct	1.0	0.0	0.90	0.09
LLaMA3-orgCP	1.0	0.0	0.50	0.49
LLaMA3-InsCP	0.99	0.01	0.01	**0.99**

Table 1: Language alignment benchmark.

model	TruthfulQA mc2 ↑		ToxiGen toxicity ↓		BOLD sentiment ↓	
language	EN	ZH	EN	ZH	EN	ZH
LLaMA3-instruct	51.6	52.7	0.10	0.14	0.54	0.55
LLaMA3-orgCP	50.8	50.5	0.12	0.26	0.61	0.68
LLaMA3-InsCP	**51.8**	**53.8**	**0.07**	0.16	0.56	**0.52**

Table 2: Reliability benchmark

5 Experimental Result

5.1 Language alignment evaluation

We present the percentage of responses among 2000 prompts generated by the models. The experimental findings are summarized in Table 1. Our observations are as follows: (1)**LLaMA3-instruct exhibits poor language alignment:** As indicated in Table 1, when provided with Traditional Chinese input prompts, LLaMA3-instruct frequently generates output in English. This lack of alignment between the input and output languages can lead to language nonalignment issues during usage. (2)**The same data used with the original CP method fails to achieve proper alignment:** A key distinction between InsCP and the original CP lies in their respective language learning capabilities. We observed that with the same data size, InsCP enables LLMs to acquire language proficiency more effectively. (3)**LLaMA3-InsCP demonstrates remarkable language proficiency:** Regardless of whether provided with English or Traditional Chinese input prompts, LLaMA3-InsCP consistently responds in the appropriate language.

5.2 Reliability evaluation

In Table 2, we present the results of the models' reliability. Our experiments were conducted in both English and Chinese to ensure that our model does not compromise its RLHF ability in either language. Across each benchmark, we observe that the orgCP model consistently achieves lower scores compared to the other models. On the other hand, LLaMA3-InsCP retain the RLHF ability, allowing it to defend against toxic inputs and generate non-harmful context during inference.

5.3 Knowledge benchmark

In Table 3, we present the scores from six knowledge benchmark tests. In Chinese-related benchmarks, we observed that the model after InsCP exhibited some improvements compared to both orgCP and the original model. These findings indicate that InsCP can effectively preserve the LLM's

model	ARC	Hellaswag	MMLU	C-eval-tw	TMMLU+	TTQA
	ACC ↑	ACC ↑	ACC ↑	ACC ↑	ACC ↑	ACC ↑
LLaMA3-instruct	60.5	81.8	67.2	47.3	43.0	23.3
LLaMA3-orgCP	57.5	81.3	66.1	48.5	41.3	41.3
LLaMA3-InsCP	**61.6**	81.7	65.6	**48.9**	**41.9**	**48.5**

Table 3: Knowledge benchmark

model	MT-Bench	
language	EN ↑	ZH ↑
LLaMA3-instruct	7.8	4.1
LLaMA3-orgCP	4.3	4.6
LLaMA3-InsCP	7.6	**6.7**

Table 4: MT-Bench

model	MT-Bench-JP
LLaMA3-instruct	4.9
LLaMA3-orgCP-JP	4.8
LLaMA3-InsCP-JP	6.6

Table 5: MT-Bench-JP

inherent abilities while also enhancing its performance in target language domains.

5.4 MT-Bench and MT-Bench-JP

In Tables 4 and 5, MT-Bench further highlights the distinctions between orgCP and InsCP. We note that outputs from orgCP often contain irrelevant text that deviates from our input prompts. Moreover, the orgCP model appears to forget how to appropriately conclude conversations. Additionally, due to the inclusion of language alignment criteria in GPT-4 evaluation, we observe a significant disparity between the InsCP model and LLaMA3-instruct. While LLaMA3-instruct predominantly responds in English for most questions, the InsCP model demonstrates the ability to discern the language input by the user. We observe a distribution similar to that of Traditional Chinese MT-Bench in Table 5 in Japanese domain.

6 Limitations of InsCP

As discussed in Section 4.1, the choice of data used in InsCP significantly influences its outcomes. Our experiments indicate that conducting InsCP necessitates the utilization of low-perplexity instruction-following data, which can be challenging to acquire in abundance for certain languages. Consequently, we opted to perform InsCP using small datasets, which we believe is a more generalizable approach for languages with limited resources. Nonetheless, both data size and data quality remain challenges when implementing InsCP.

7 Conclusion

In this work, we introduce a novel pipeline called InsCP designed to facilitate the transfer of LLMs into non-English domains. Through InsCP, LLMs can retain their inherent abilities while also acquiring the capability for language alignment in the target language and gaining knowledge of the target domain. Additionally, we demonstrate that InsCP does not necessitate extensive data, thereby consuming fewer resources and less time. Remarkably, even with a small amount of data, InsCP can transform English-based LLMs into models aligned with the target language, a stark contrast to the resource-intensive traditional pipeline. InsCP paves the way for future LLMs, primarily fine-tuned in specific languages, to swiftly transfer their abilities to other languages.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

AI@Meta. 2024. Llama 3 model card.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to re-warm your model? In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da shan Shiu. 2023. Advancing the evaluation of traditional chinese language models: Towards a comprehensive benchmark suite.

Shih-Cheng Huang, Pin-Zu Li, Yu-Chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tzong-Han Tsai, and Hung yi Lee. 2024. Chat vector: A simple approach to equip llms with instruction following and model alignment in new languages.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2022. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*.

Chen-An Li and Hung-Yi Lee. 2024. Examining forgetting in continual pre-training of aligned large language models.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Yen-Ting Lin and Yun-Nung Chen. 2023. Language models for taiwanese culture. Code and models available at https://github.com/MiuLab/Taiwan-LLaMa.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to!

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Zhi-Rui Tam, Ya-Ting Pai, Yen-Wei Lee, Sega Cheng, and Hong-Han Shuai. 2024. An improved traditional chinese evaluation suite for foundation model.

Gemini Team, Rohan Anil, Sebastian Borgeaud, and Jean-Baptiste Alayrac et al. 2024. Gemini: A family of highly capable multimodal models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and Matthias Gallé *et al.* 2023. Bloom: A 176b-parameter open-access multilingual language model.

YuLan-Team. 2023. Yulan-chat: An open-source bilingual chatbot. https://github.com/RUC-GSAI/YuLan-Chat.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Appendix

A.1 LLaMA3-instruct chat template

To initiate a completion with LLaMA3-instruct, one must adhere to the following format:

```
<| begin_of_text |>
<| start_header_id |>user<|end_header_id|>
{{ inputs }}<|eot_id|>
<| start_header_id |> assistant <|end_header_id
    |>
{{model_response}}
```

The InsCP template is shown below:

```
<| begin_of_text |>
<| start_header_id |>user<|end_header_id|><|
    eot_id|>
<| start_header_id |> assistant <|end_header_id
    |>
{{InsCP_data}}<|eot_id|>}
```

A.2 Training Detail

We utilize LLaMA3-instruct as our base model, and both the original continual pre-training and instruction continual pre-training are configured with the following hyperparameters: a learning rate of 3e-5, AdamW optimizer with beta1 of 0.9 and beta2 of 0.95, batch size set to 1 per device (utilizing 64 GPUs), and training conducted for 10 epochs.

A.3 Generation Strategy

We employ vLLM as our generation tool, incorporating LLaMA3's system prompt in each generation to harness the full potential of the LLM. For vLLM, we set the following generation parameters: maximum tokens to 1024, temperature to 0.8, top-p sampling to 0.9, and seed fixed at 42 to facilitate result reproducibility. Additionally, we maintain default values for other generation configurations in vLLM.

A.4 MT-Bench evaluation prompt

In the Traditional Chinese MT-Bench, we predominantly adhere to the evaluation prompts provided by the authors. However, to delve deeper into testing the LLM's language alignment ability, we introduce an additional prompt in Traditional Chinese: "If the assistant's answer is in a language other than Traditional Chinese, please give it a score of 0." This prompt instructs GPT-4 to assign a score of 0 to responses that are not in the correct language, thereby enabling a more rigorous assessment of language alignment capabilities. For Japanese MT-Bench, we also add the prompt in Japanese: "If the assistant's answer is in a language other than Japanese, please give it a score of 0.", in order to meet the language alignment requirement we want to obseve.

Analyzing the Linguistic Priors of Language Models with Synthetic Languages

Alessio Tosolini[1,2] **Terra Blevins**[3,4]
[1]Yale University
[2]Paul G. Allen School of Computer Science, University of Washington, Seattle
[3]Faculty of Computer Science, University of Vienna
[4]Khoury College of Computer Sciences, Northeastern University
Correspondence:, alessio.tosolini@yale.edu, t.blevins@northeastern.edu

Abstract

While modern language model architectures are often assumed to be language-agnostic, there is limited evidence as to whether these models actually process the vast diversity of natural languages equally well. We investigate this question by analyzing how well LMs learn carefully constructed artificial languages containing a variety of verbal complexity, ranging from simple paradigms to covering far more verb classes than occur in natural languages. Rather than learning all languages equally efficiently, models trained on these languages show strict preferences for processing simpler languages. Furthermore, while *some* observed model preferences mimic human linguistic priors, we find that they correspond to model memorization of its training data rather than generalization from it. This finding suggests that while model behavior often mimics human language understanding, the underlying causes of their proficiencies are likely very different.

1 Introduction

Transformer-based language models (LMs) are often assumed to be language-agnostic, or to learn all natural languages equally well. This has lead to their widespread use for different languages (Scheible et al., 2024; Ahmed et al., 2024, i.a.) and multilingual modeling (e.g., Üstün et al., 2024).

However, there is immense linguistic diversity in the world's languages, and human learners acquire aspects of these languages at different rates. For example, children take longer to learn the opaque Dutch gender system, mastering it by age six (Tsimpli, 2014), while children master the transparent Spanish gender system by three and a half, if not sooner (Lew-Williams and Fernald, 2007). It remains an open question as to whether this complexity similarly affects model acquisition of different languages: previous work exploring the differences in language modeling capabilities presents mixed results on the effect of morphological complexity on language modeling (Cotterell et al., 2018; Mielke et al., 2019; Park et al., 2021; Arnett and Bergen, 2024), and typological differences can impact the performance of models intended to be language-agnostic (Gerz et al., 2018). Furthermore, there is limited evidence whether LMs are even constrained to learning natural linguistic phenomena as humans are (Kallini et al., 2024).

We address this question by testing if LMs demonstrate human-like learning patterns when acquiring new, artificial languages. Specifically, we ask: **Do LMs exhibit linguistic priors favoring certain conjugation paradigms over others?**. We center our behavioral analysis on a single grammatical feature—verb conjugation—in a wide variety of linguistically plausible and implausible settings as a controlled case study into the effect of linguistic grammatical complexity on transformer-based modeling of language.

To evaluate LMs for these linguistic priors, we first construct artificial languages using a probabilistic context-free grammar (PCFG). These languages cover a wide range of (plausible and implausible) conjugation complexity while controlling for other confounding variables found in natural languages. We then test how proficiently and efficiently language models learn these languages by measuring their mastery of both subject-verb agreement (a commonly used linguistic test for LMs, see Gulordava et al., 2018), as well as a novel behavioral experiment for *verb class identification* in these languages throughout the training process.

Our experiments find that language models acquire more complex languages (i.e., those with more verb classes) more slowly. However, they achieve close to 100% accuracy on seen verbs given enough data, even in cases where the number of verb classes is far larger than naturally occurs in human languages. The models also perform significantly worse on novel verbs than those seen during

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 7–15
August 1, 2025 ©2025 Association for Computational Linguistics

training, with the performance degradation increasing with the number of verb classes; this indicates that these models do not learn to generalize from the standard conjugation patterns shown to them during pretraining.

These findings suggest both that (1) these models are *not* language-agnostic, but are instead sensitive to the complexity of the target language, and that (2) behavior that resembles human-like language learning in models may actually be *memorization* of the training data, rather than *generalization* to the underlying linguistic rules. Put another way, correlations between model and human behavior do not necessarily indicate that their underlying mechanisms are the same. In light of these findings, we recommend future work analyzing model language learning to incorporate evaluations that disentangle these factors when probing language model behavior.

2 Methodology

This section presents our method for generating artificial languages with the desired characteristics (Section 2.1) and our behavioral experimental setting to test model proficiency on subject-verb agreement in these languages (Section 2.2).

2.1 Artificial Language Generation

To evaluate how well LMs can learn languages across different verb settings, we generate artificial languages with the desired features using a Probabilistic Context-Free Grammar (PCFG), an extension of context-free grammars that assigns probabilities to transitions between states, allowing for the stochastic generation of sentences. We focus our analysis on these artificial languages to control for various confounding factors found in natural languages, including but not limited to semantics, irregularities, ambiguity, and dialectal variation, that make direct comparisons difficult.

We define our PCFG with a set of parameters describing the language's word formation, syntax, and inflectional rules. For verb paradigms, this parameterization allows us to perform controlled ablations across various experimental settings. Specifically, for our experiments, we generate ten languages for each of the {1, 2, 3, 5, 8, 16, 32, 64} verb class settings and report the average performance and standard error in a given setting. There is no overlap in the suffixes between any two verb classes, and verb paradigms are fully regular.

Other parameterization of our PCFG is informed by common natural distributions of language features to ensure our artificial languages are as similar to natural ones as possible. In each language, the number of roots generated per part-of-speech approximates 1% of English senses in Kaikki (Ylönen, 2022), with nouns approximating 0.5% of senses since jargon is often overrepresented in nouns (Table 1). As Zipf's Law is ubiquitous in human language at many scales (Williams et al., 2015), the distribution from which words are selected is drawn from a Zipfian distribution. A skew of 1.2 is used for our word distribution, based on the empirical distribution found in the American National Corpus (Piantadosi, 2014). The verb class assigned to a verb is similarly drawn from a Zipfian distribution with a skew of 1.

We also allow for features (such as nominative for subjects) to be passed between states in the PCFG during generation (Figure 1); this enforces subject-verb agreement on person and number features within each sentence. A more detailed explanation of creating the artificial languages and generating sentences is given in Appendix A.

Part of Speech	Items	Kaikki Senses
Adjective	2000	199759
Determiner	1	387
Noun	4000	856855
Preposition	15	1337
Pronoun	6	1053
Verb	2000	220457

Table 1: Word counts per part of speech in our artificial languages versus Kaikki sense counts for English.

2.2 Model Training and Evaluation

When training language models on our artificial languages, we consider three factors: the verb

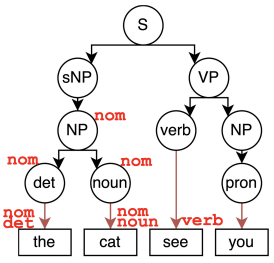
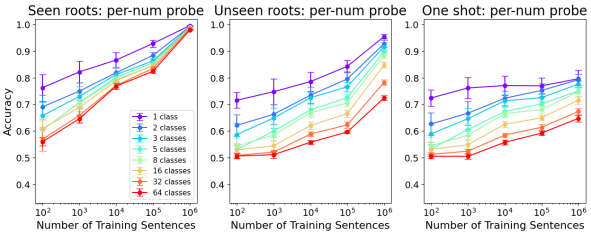
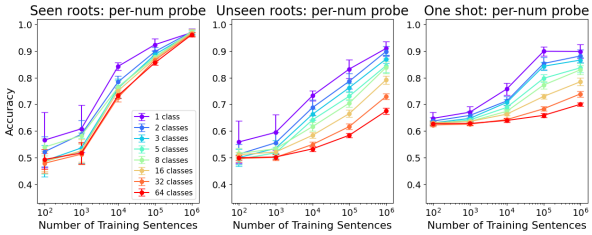


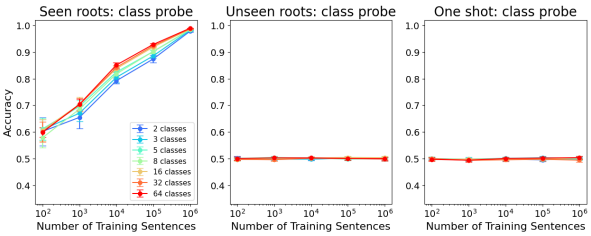
Figure 1: Sample English PCFG. The nominative feature nom passes from the child of the subject noun phrase *sNP* to its descendants, allowing for subject-verb agreement to be enforced later in the generation pipeline.



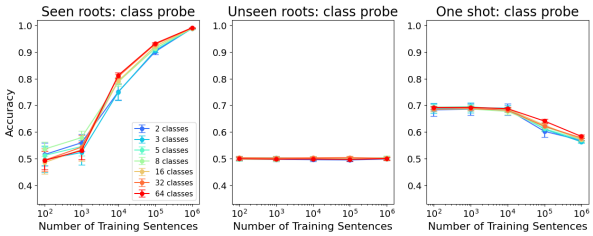
(a) Per-num behavioral test on GPT-2 trained from scratch



(b) Per-num behavioral test on finetuned GPT-2



(c) Class behavioral test on GPT-2 trained from scratch



(d) Class behavioral test on finetuned GPT-2

Figure 2: Behavioral test results for models trained on the generated languages at different training data sizes.

paradigm (or number of verb classes in a language), the size of the training dataset, and the model training scheme. We construct training datasets of $10^{2,3,4,5,6}$ grammatical training sentences from each of the generated languages; the two training schemes considered are training a randomly initialized GPT-2 model from scratch versus finetuning the pretrained GPT-2 model for English (Radford et al., 2019), both of which have approximately 124 million parameters. We train LMs on each combination of these settings (and across all ten languages per verb paradigm).

We then evaluate these models on how well they learn the artificial languages they are trained on by testing whether they can distinguish grammatical examples of the language from ungrammatical ones.[1] Specifically, we construct evaluation sets consisting of minimal pairs of grammatical and ungrammatical sentences and measure the perplexity of the model on each sentence; models that are well-fit to the training language should prefer (achieve a lower perplexity on) the grammatical sentences.

We consider two types of behavioral tests to probe how well the models learn subject-verb agreement and the verb classes. These tests determine the error type shown during inference in a minimally different pair of (grammatical, ungrammatical) sentences: the *per-num* test (Case 1), where the verb in the ungrammatical sentence takes a suffix marking a different person and/or number feature for that class:

(1) a. El perro *escucha* el gato.
"The dog *hears* the cat."
b. *El perro *escucho* el gato.
"The dog *hear* the cat."

and the *class* test (Case 2), where the ungrammatical verb takes a suffix from a different verb class agreeing with the subject's person and number:

(2) a. El perro *escucha* el gato.
"The dog *hears* the cat."
b. *El perro *escuche* el gato.
"The dog *hears*[2] the cat."

The evaluation sets contain 5,000 sentence pairs; we define accuracy as the percentage of test sentence pairs where the grammatically correct sentence's perplexity is less than that of the ungrammatical sentence. Thus, we measure the cases where the model assigns a higher likelihood to the grammatical case as a proxy for how well it models conjugation in the generated languages.

Testing covers three settings, varying whether test verb roots are seen during training: *seen roots*, where the model is evaluated on verb roots from the training; *unseen roots*, which evaluates the model on held-out verb roots to test model generalization; and *one-shot*, where the model is given one demonstration using a hitherto unseen verb before being

[1] This is a common approach for surfacing linguistic knowledge in LMs (e.g., Liu et al., 2019), particularly in the case of subject-verb agreement (Gulordava et al., 2018).

[2] Note that there is no equivalent, ungrammatical English translation for the *class* test, as English does not have verb classes that correspond to multiple regular conjugation paradigms like in Spanish.

tested on that same verb root. Given the set s of LMs trained across each (data scale, verb paradigm, and training scheme) combination, we evaluate s on all described (i) test types and (ii) evaluation settings. We report the mean performance and standard error across the ten runs for each combination.

3 Results

This section presents our language learning experiments. We find that while LMs model simpler inflectional paradigms more easily (indicating that they are not agnostic to language complexity), they struggle to generalize to new verbs across experimental settings, including on linguistically plausible verbal paradigms found in natural languages.

3.1 Per-Num Agreement Evaluation

Figures 2a and 2b show *per-num* behavioral test outcomes (where negative samples contain incorrect subject-verb agreement) on models trained from scratch and finetuned GPT-2, respectively. Across settings, adding verb classes to the generated languages generally corresponds to worse performance on (and slower acquisition of) subject-verb agreement by LMs.

For seen roots, all models achieve high accuracies of 97.5% or greater at the largest data size (1M training sentences). However, acquisition time varies across model and verb class settings: languages with more verb classes consistently need more data to achieve comparable accuracies to those with fewer classes[3]. Pretrained GPT-2 also learns to prefer correctly conjugated seen verbs *slower* than models trained from scratch.

Unsurprisingly, agreement accuracy on unseen roots is lower than on seen roots across comparable experiments.[4] However, we see the same relative trends here as on seen roots: more data improves conjugation accuracy (though now with larger gaps between the best- and worst-performing LMs), and finetuned GPT-2 continues to underperform in limited data settings. For unseen verbs, though, the performance gap between the randomly initialized and finetuned LMs is smaller, particularly on languages with eight or more verb classes.

Finally, we find that providing the model with one correctly conjugated demonstration does not consistently improve accuracy over the unseen verb ("zero-shot") setting. In many cases, the models perform similarly in both settings, and high-data regimes often perform *worse* when given a correct example. This, in addition to the unseen verb results, suggests the models do not learn abstract conjugation patterns when trained on these languages.

3.2 Verb Class Evaluation

Figure 2c presents the *class* behavioral test (where negative samples contain a verb that is correctly conjugated, but with the wrong class pattern) results on models trained from scratch; Figure 2d shows the corresponding results for finetuned GPT-2. Unsurprisingly, we observe random chance performance (50%) on unseen verbs for both the randomly initialized and finetuned models–as the models cannot predict the correct class for verbs not seen during training.

More surprisingly, randomly initialized models are also unable to outperform random chance in the one-shot setting, suggesting that these models can not generalize knowledge about underlying verb classes during inference. While one-shot evaluations of the finetuned model outperform this in low-data settings (achieving $\sim 68\%$ accuracy), this is roughly what would occur if the model always chooses sentences where the prompt and test verb are identical (occurring $\sim \frac{1}{6}$ of the time across conjugations), and chooses randomly otherwise; this performance also occurs on the *per-num* test (Figure 2b). Thus, this behavior is likely caused by the pretrained GPT-2 exhibiting a strong copying preference (Olsson et al., 2022), but not generalizing beyond that.

On seen verbs, model performance again generally improves with more data, but we see smaller performance gaps across languages with different verb class counts, particularly at smaller data scales. Furthermore, model accuracy with more verb classes tends to be *higher* than those with fewer classes, though with more variation than observed with *per-num* probing. The discussion offers a possible hypothesis for this phenomenon.

4 Discussion

This paper investigates whether LMs exhibit linguistic priors for natural and unnatural conjugation paradigms. Our probing experiments find that LMs are much more efficient at modeling person-number agreement for languages with simpler verb

[3]E.g., Training from scratch on 100 sentences with one verb class achieves a mean accuracy of 76.2%, while it requires 10k sentences to get a similar accuracy over 64 classes.

[4]Limited generalization has been observed for other linguistic tasks in transformers (Liu and Hulden, 2022).

paradigms, mirroring human learning of languages. They also corroborate prior work indicating that neural LMs prefer human languages to unnatural ones (Alamia et al., 2020; Kallini et al., 2024). However, this primarily holds for verbs seen during training; models perform much worse at judging subject-verb agreement on novel verb roots, in contrast with the strong generalization shown by human speakers on this task (e.g. Berko, 1958).

Furthermore, we find that LMs adopt unnatural verb paradigms[5] almost as well, given enough training data. This result, in conjunction with degraded performance exhibited on unseen verbs, indicates that model learning of the generated languages is likely heavily dependent on **memorization** rather than **generalization** of the training data, particularly in the *class* behavioral test setting. While this trade-off has been documented in LMs for downstream NLP tasks (Tänzer et al., 2022; Zheng and Jiang, 2022), we find that it also affects the model when learning lower-level linguistic knowledge.

Even more unnaturally, models trained on languages with more complex paradigms are slightly *better* at identifying correct verb classes, with the best performance occurring on 32 and 64 classes. –far beyond what appears in most natural languages. We hypothesize that this behavior is due to how models and their inputs are parameterized: as the number of classes increases, the set of verb roots a suffix can follow (according to the training data) becomes smaller, allowing the model to be more confident about the bigram's conditional probability. However, this finding contrasts sharply with human language learning, where many unrelated paradigms are typologically improbable due to the unreasonable amount of memorization required for humans to model them correctly.

Based on these results, we argue that while language model learning of verbal paradigms may resemble human learning, the underlying mechanisms driving these behaviors are likely very different. Future work comparing model behavior with humans should control for these similarities by also looking at the underlying mechanisms driving model performance.

5 Limitations

Using carefully constructed artificial languages allows us to isolate syntactic complexity's effects on language learnability and to consider a broad, systematic complexity distribution. However, this means that these languages are not natural (particularly regarding the absence of semantics), which limits the findings presented here. Future work should replicate these experiments in a more natural setting to verify that our findings remain valid in such conditions.

Another limitation of this work is the size of the language models: computational limitations and the number of models considered in our experiments (800 trained LMs across experimental settings) limited the model size considered to one setting, GPT-2 Small (124M parameters). Finally, there are many aspects of complexity in natural language, with the number of verb classes being just one aspect. Whether our findings hold for other linguistic phenomena, such as noun classes (i.e. gender), freedom in word order, degree of syncretism, morphophonological alternations, etc. remains an open area for future research.

[5]I.e., more verb classes than in most natural languages.

Acknowledgments

We would like to thank Luke Zettlemoyer for feedback on early stages of this work. We would additionally like to thank Claire Bowern for access to her lab compute resources for model training in the later stages of this work.

References

Murtadha Ahmed, Saghir Alfasly, Bo Wen, Jamal Addeen, Mohammed Ahmed, and Yunfeng Liu. 2024. AlclaM: Arabic dialect language model. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 153–159, Bangkok, Thailand. Association for Computational Linguistics.

Andrea Alamia, Victor Gauducheau, Dimitri Paisios, and Rufin VanRullen. 2020. Comparing feedforward and recurrent neural network architectures with human behavior in artificial grammar learning. *Scientific Reports*, 10(1):22172. Publisher: Nature Publishing Group.

Catherine Arnett and Benjamin K Bergen. 2024. Why do language models perform worse for morphologically complex languages? *arXiv preprint arXiv:2411.14198*.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 2 (Short Papers), pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Casey Lew-Williams and Anne Fernald. 2007. Young children learning spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, 18(3):193–198. PMID: 17444909.

Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21:1112 – 1130.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Raphael Scheible, Johann Frei, Fabian Thomczyk, Henry He, Patric Tippmann, Jochen Knaus, Victor Jaravine, Frank Kramer, and Martin Boeker. 2024. GottBERT: a pure German language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21237–21250, Miami, Florida, USA. Association for Computational Linguistics.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus generalisation in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7578.

Ianthi Tsimpli. 2014. Early, late or very late: Timing acquisition and bilingualism: A reply to peer commentaries. *Linguistic Approaches to Bilingualism*, 4.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Jake Ryland Williams, Paul R. Lessard, Suma Desu, Eric Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Preprint*, arXiv:1406.5181.

Tatu Ylönen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 1317–1325. European Language Resources Association.

Xiaosen Zheng and Jing Jiang. 2022. An empirical study of memorization in nlp. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6265–6278.

A Methodological Details

A.1 Artificial Language Generation

An overview of the pipeline used to generate sentences is described in Figure 3.

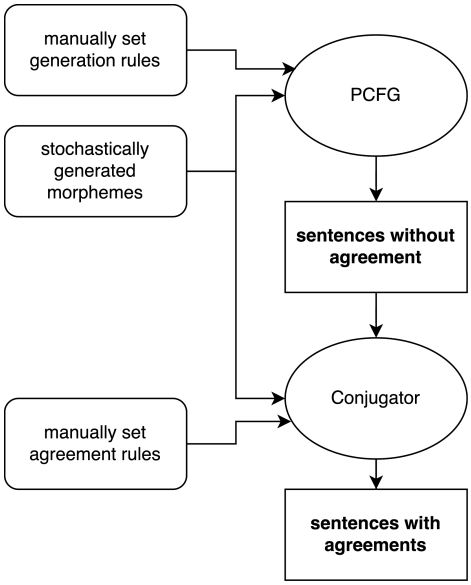


Figure 3: Overview of Sentence Generation with Artificial Languages.

A.1.1 Morpheme Generation

The parts of speech and number of morphemes that belong in each category are set manually. For this experiment, about 8000 morphemes were used in total. The number of morphemes for each part of speech was chosen to approximate 1% of the number of Kaikki's senses for each of English's part of speech (Ylönen, 2022). Approximately 0.5% of nouns were used instead of 1% since nouns disproportionately have technical vocabulary or jargon in Kaikki's database which we do not care to replicate. Additionally, only one determiner, which inflects for number, and six pronouns for each combination of number (singular, plural) and person (1st, 2nd, 3rd) were chosen to simplify conjugation paradigms.

A.1.2 Probabilistic Context Free Grammar

Probabilistic context-free grammars (PCFG) are an extension of a context-free grammar where each input state's production rules take a probability. Since the final result of a PCFG resembles a syntactic tree, it allows us to create sentences with customizable linguistic structures. Pseudocode for a simplified English grammar is provided in Figure 4. All generations start with an "S" state. Generation rules may apply categorically, such as rule (1), which determines that sentences produce a subject noun phrase followed by a verb phrase with probability 1. Other states may have two or more possible outcomes, as demonstrated by rule (3), which determines that noun phrases may produce a determiner followed by a noun or a pronoun with equal probability. All preterminal states are lowercase, while non-preterminal and non-terminal states are required to have at least one capital letter. This simplified grammar does not include adjectives or prepositions and thus is incapable of handling recursion, but the full grammar for the artificial languages does.

```
generation_rules = [
    S → [sNP, VP], 1
    sNP → [NP.nom], 1
    NP → [det, noun], 0.5, [pron], 0.5
    VP → [verb, NP], 0.7, [verb], 0.3
]
```

Figure 4: Generation Rules for a simplified grammar.

In order to handle conjugation, states are assigned features, as demonstrated by rule (2). Not demonstrated is the deletion of a feature and the addition of a tag feature that allows for long-distance agreement. Unless explicitly stated by a generation rule, a state passes all of its features to its children. This can be seen in Figure 1, where the subject NP with feature `nom` (short for nominative, i.e., subject of a sentence) passes on the feature to all of its children states.

Preterminal nodes are represented by the part of speech that will beget a morpheme from the vocabulary. A simplified vocabulary can be seen in Figure 5. In order to mimic the naturalistic distribution of words in human languages, generation rules for preterminal states function differently from the rest of the PCFG. The morpheme that is chosen by the preterminal state is chosen according to a Zipfian distribution with skew = 1.2. Additionally, the part of speech of a terminal node is added to its set of features.

Rules dictating universal features of certain parts of speech may be added at this step as well. For example, in our simplified toy vocabulary, all nouns are assumed to take the feature `3rd`. This is not included in the toy grammar's rules, but will be essential in determining which terminal states agree with which other terminal states.

```
vocabulary = {
    det: [“the”],
    noun: [“cat”, “dog”],
    verb: [“see”, “miss”],
    pron: [“you”]
}
```

Figure 5: Vocabulary for a simplified grammar. The morphemes in this example are not stochastically generated for a clearer example.

A.1.3 Conjugator

The conjugator works by (i) determining which morphemes agree with other morphemes and (ii) applying each inflection rule to each word in the sentence that has a given feature set.

After the PCFG generates a sentence without agreement, the first step is determining which lemmas agree with which other lemmas. We define agreement as a terminal state *copying* a state from another terminal state according to rules set by the user. For example, as demonstrated by Figure 6, we can see our toy grammar's verbs are required to copy one feature relating to number and one feature relating to person from the nominative noun constituent, as seen by the definition of `agreement_rules`. In the toy example in Figure 1, the verb *see* copies the feature `3rd` and `sg` (applied through rules not shown in previous figures) from *cat* to have the feature set `verb, 3rd, sg`. Note that if any word seeking agreement finds more than one word to agree with, or is unable to find exactly one of each of the features it aims to copy, generation fails.

The second step is applying inflections to the sentence, now that all words requiring conjugation have copied features from the word that they are agreeing with. Inflections apply to any word that changes form based on some features. For example, as demonstrated by Figure 6, we can see our toy grammar's verbs take a suffix *-s* iff it has features `3rd` and `sg` or otherwise it does not take any inflections, as seen by the definition of `conjugations`. Conjugations may also apply to words which did not gain features from the agreement rules, such as nouns pluralizing with the suffix *-s* if it gained the `pl` feature from a generation rule.

We generate datasets on 8 different numbers of verb classes: {1, 2, 3, 5, 8, 16, 32, 64}. In all datasets, any verb agrees with the subject of the sentence in person and number, for a total of 6

```
agreement_rules = [
    # Verbs agree with the nom nouny word
    # Verbs must then copy the word's number
    # Similarly, they then copy the person
    {"verb": [["nom", "nouny"],
              [["sg", "pl"],
               ["1st", "2nd", "3rd"]]]}
]
conjugations = [
    ["verb", {
       “-s”: 3rd.sg,
       “-”: otherwise
    }],
    ["noun", {
       “-s”: pl,
       “-”: otherwise
    }]
]
```

Figure 6: Pseudocode for agreement rules and conjugations for a simplified grammar. Nouny is defined as being either the feature `pron` or `noun`.

possible suffixes given a verb root (3 person x 2 number). In datasets with n verb classes where $n > 1$, verbs are assigned to one of n classes. Each of these classes has a unique set of suffixes for each combination of person and number for subject-verb agreement, for a total of $6n$ verbal suffixes per language.

B Replicability Details and Miscellanea

This section provides additional details on our experimental setting for documentation and replicability purposes.

The language models trained in these experiments have the GPT-2 Small architecture with 124.4M trainable parameters. We consider both a randomly initialized version of this architecture and the pretrained GPT-2 hosted through Huggingface,[6] which is released under the MIT license; intended use of this artifact beyond the license is not clearly stated.

In our experiments, we trained 800 models (10 runs, 8 languages per run, 5 models with varying training data amounts per language, 2 base models - English pretrained vs. randomly initialized), and the training dataset sizes ranged from 100 sentences to one million sentences; we evaluated each model

[6] https://huggingface.co/openai-community/gpt2, as accessed before and on 02/15/2025.

across six settings (two behavioral tests, three root settings) with 5,000 sentences each; though, models trained on languages with one verb class were not evaluated on the *class* behavioral test. There is no overlap in sentences from the training and test set. We use 10 GPUs for both training and evaluation, one for each run of 80 models. Training and evaluating 80 models took approximately one day per GPU. Given our training batch size of 1 and single training epoch, this corresponds to between 100 to 1,000,000 training passes per model.

Unstable Grounds for Beautiful Trees? Testing the Robustness of Concept Translations in the Compilation of Multilingual Wordlists

David Snee, Luca Ciucci, Arne Rubehn, Kellen Parker van Dam, Johann-Mattis List

Chair for Multilingual Computational Linguistics, University of Passau, Passau, Germany

Abstract

Multilingual wordlists play a crucial role in comparative linguistics. While many studies have been carried out to test the power of computational methods for language subgrouping or divergence time estimation, few studies have put the data upon which these studies are based to a rigorous test. Here, we conduct a first experiment that tests the robustness of concept translation as an integral part of the compilation of multilingual wordlists. Investigating the variation in concept translations in independently compiled wordlists from 10 dataset pairs covering 9 different language families, we find that on average, only 83% of all translations yield the same word form, while identical forms in terms of phonetic transcriptions can only be found in 23% of all cases. Our findings can prove important when trying to assess the uncertainty of phylogenetic studies and the conclusions derived from them.

1 Introduction

While the quantitative turn in historical linguistics has been met with a considerable amount of skepticism for a long time (Holm, 2007; Geisler and List, 2022), phylogenetic methods – originally developed to infer phylogenies of biological species – have by now become the new state of the art in the field, replacing more traditional methods for subgrouping almost completely. Given that most linguistic approaches to phylogenetic reconstruction make use of lexical data, *multilingual wordlists* play a crucial role in phylogenetic reconstruction in historical linguistics.

The compilation of multilingual wordlists itself is quite tedious. Starting from a list of concepts, scholars must translate the concepts into all target languages under investigation. The translation into the target languages, however, is no standardized procedure, but may require various steps, including the consultation of informants, the consultation of published resources, or the inspection of archived material. In all these cases, scholars who compile a wordlist must weight their evidence carefully, in order to avoid errors. Given the complexity of this process, it is no surprise that errors can easily slip into the translations. A given concept may lack a direct translational equivalent in a given language, or there may be several good candidates from which scholars must select the most appropriate ones. As a result, there is a great risk that multilingual wordlists compiled for phylogenetic studies show a considerable amount of idiosyncrasies that might have an impact on the phylogenies scholars compute from them.

Studies that try to measure the amount of inconsistency in multilingual wordlists – introduced by the translation of concepts into target languages – are lacking so far. Here, we present a first attempt to shed light on the robustness of the concept translation task, taking advantage of the fact that recent efforts have produced large-scale repositories of standardized multilingual wordlists (List et al., 2022). In the following, we will give a short overview on previous discussions and studies that focus on the translation of concepts in multilingual wordlist compilation (§ 2). After this, we introduce the materials and methods by which we try to evaluate the robustness of concept translation (§ 3). Having presented the results (§ 4), we discuss them in more detail and share some ideas to improve the enterprise of wordlist compilation in historical linguistics (§ 5).

2 Background

Phylogenetic approaches rely on multilingual wordlists compiled through lexicostatistic methods, dating back to Swadesh's foundational work (Swadesh, 1950, 1952, 1955). A multilingual wordlist in this context is a list of concepts translated into one or more target languages (List, 2014).

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 16–28
August 1, 2025 ©2025 Association for Computational Linguistics

Although typically emphasizing their difference with respect to Swadesh's lexicostatistics, modern phylogenetic approaches all build on this onomasiological (i.e. *concept-based*) approach that takes the concept as the major aspect by which languages are compared. Building on Geisler and List (2010), we can identify five major steps in the typical workflow applied in modern approaches to phylogenetic reconstruction. Starting from the compilation of a concept list (1: *concept list compilation*), the concepts are translated into the target languages (2: *concept translation*) in order to create an initial *comparative wordlist*. This wordlist is then used as the basis for the identification of cognate words (3: *cognate identification*). Having converted the information on cognate words into a numerical or computer-readable format (4: *cognate coding*), scholars then employ their phylogenetic method of choice in order to compute a phylogeny of the languages in question (5: *phylogenetic reconstruction*).

Up to now, most critics of lexicostatistics and its modern equivalents have concentrated on either the stage of cognate identification or the resulting phylogenetic methods. Cognate identification is often criticized as being flawed due to undetected borrowings (Donohue et al., 2012). When disputing over phylogenetic methods, there have been long-standing debates about the complexity of the models employed, as reflected in the debate about the age of Indo-European, where models differing in complexity yielded quite different age estimates (Bouckaert et al., 2012; Chang et al., 2015; Kassian et al., 2021; Heggarty et al., 2023).

What has much less often been discussed in the context of phylogenetic methods, however, are the first two stages of the workflow, that is, the stage of concept list compilation, and the stage of concept translation. While it has been clear for a long time that different concept lists often yield different phylogenies (Chén, 1996; McMahon et al., 2005), a closer discussion regarding the impact of concept lists on the results of phylogenetic analyses has not been carried out so far. The same holds for the translation of concepts into target languages. While Geisler and List (2010) found that concept translation across Romance languages in two independently compiled multilingual wordlists differs by about 10%, and List (2018) and Häuser et al. (2024) could show that selecting but one out of several translations for the same concept in the same language can have direct consequences on the resulting phylogenies, no closer investigation regarding the degree of variation in concept translation or the impact of concept translation on phylogenetic analyses has been conducted up to now.

3 Materials and Methods

3.1 Materials

In order to investigate variation in word choice across multilingual wordlists, it is important to find wordlists that have been compiled independently for the same language varieties, containing at least a certain subset of identical concepts. In order to identify such data, we checked datasets published as part of the Lexibank repository (https://lexibank.clld.org, List et al. 2022), searching specifically for those cases where several languages from the same language family or subgroup are available in the form of multilingual wordlists created by different authors.

Lexibank uses Cross-Linguistic Data Formats (CLDF, https://cldf.clld.org, Forkel et al. 2018) to standardize multilingual wordlists along the three dimensions of language, meaning, and form. Languages are linked to Glottolog (https://glottolog.org, Hammarström et al. 2024) in order to ensure that languages can be easily identified across sources, even if they are given different names in the original datasets. Concepts are mapped to Concepticon (https://concepticon.clld.org, List et al. 2025a), a reference catalog for semantic glosses used to elicit concepts in concept lists. This facilitates the aggregation of wordlists from different sources to allow the identification of common concepts for which different wordlists provide translations in their target languages. Phonetic transcriptions in Lexibank are unified with the help of the Cross-Linguistic Transcription Systems reference catalogue (CLTS, https://clts.clld.org, List et al. 2024), a standardized subset of the International Phonetic Alphabet that has a generative component by which detailed transcriptions of more than 8,000 speech sounds can be created and compared (see Anderson et al. 2018 and Rubehn et al. 2024).

Checking the Lexibank data in the most recent version of the repository (2.0, Blum et al. 2025a), we identified 10 groups of languages corresponding to 9 different languages families, in which two and more multilingual wordlists from different datasets could be compared. The 10 groups stem from 18 different datasets, with two

Group	Dataset A	Dataset B	Concepts			Languages			Synonymy	
			A	B	A∩B	A	B	A∩B	A	B
Bai	allenbai (Allen, 2024)	wangbai (Wang, 2024)	499	412	208	9	10	2	1.01	1.02
Chadic	kraftchadic (Kraft, 2024)	gravinachadic (Gravina, 2024)	429	717	325	67	48	5	1.04	1.10
Chinese	beidasinitic (Beijing University, 2024)	liusinitic (Lìlǐ et al., 2024)	738	202	102	18	19	12	1.18	1.17
Dravidian	dravlex (Kolipakam, 2024)	northeuralex (Dellert, 2024)	100	954	94	20	107	4	1.39	1.02
Indo-European	iecor (Heggarty et al., 2024)	starostinpie (Starostin, 2024)	170	110	88	160	19	15	1.01	1.05
Japonic	leejaponic (Lee and Hasegawa, 2024)	robbeetstriangulation (Robbeets, 2025)	210	254	152	59	101	6	1.01	1.03
Koreanic	leekoreanic (Lee, 2024)	robbeetstriangulation (Robbeets, 2025)	246	254	175	15	101	13	1.01	1.01
Tupian	galuciotupi (Galucio et al., 2024)	gerarditupi (Ferraz Gerardi and Reichert, 2024)	100	242	70	23	38	5	1.02	1.00
Uralic	northeuralex (Dellert, 2024)	syrjaenenuralic (Syrjänen et al., 2024)	954	173	147	107	7	5	1.12	1.17
Uto-Aztecan	utoaztecan (Greenhill et al., 2025)	davletshinaztecan (Davletshin, 2024)	121	100	92	46	9	3	1.22	1.00

Table 1: Selected language groups along with their original datasets and additional statistics employed in this study. References to the datasets follow the most recent publication of the data as part of the Lexibank repository. Information on the original studies in which the data were published for the first time are provided by the more recent standardized editions. The table lists the number of concepts and languages (along with the intersection), as well as the synonymy (measured by dividing the number of words by the number of concepts).

datasets (NorthEuralex, see Dellert et al. 2020, and RobbeetsTriangulation, see Robbeets et al. 2021) offering two groups each. From these 10 wordlist pairs, each consisting of two multilingual wordlists covering at least two language varieties, we manually selected 70 language pairs, making sure that all pairs represent identical languages to the best of our knowledge. Table 1 provides an overview on the 10 groups of language pairs that we compiled for this study, along with the number of matching concepts, matching language pairs, and synonymy statistics on the wordlists. Figure 1 shows the geographical distribution of the 70 languages in our sample.

3.2 Wordlist Comparison

We compare wordlists in terms of their (1) matching Glottocodes, (2) manually selected language pairs, and (3) matching concepts. Paired language varieties within each wordlist comparison are initially identified based on matching Glottocodes. This results in a total of 75 Glottocode matches across all language families to be compared. Upon closer examination of the data, it becomes clear that Glottocode comparisons alone do not allow for the comparison of identical language varieties as different subvarieties are at times encoded with the same Glottocode both inside the same dataset and in separate datasets. To quantify the effect of this discrepancy, language pairs are also manually selected based on consultation with each dataset's metadata, resulting in a total of 70 language pairs for analysis. Language pairs are then examined based on matching concepts, which are identified with the help of the Concepticon mappings provided by Lexibank.

3.3 Comparing Concept Translations

Since phonetic transcriptions in Lexibank's datasets are unified, following the system recommended by the CLTS reference catalog, one might expect that differences in concept translation can be simply identified by comparing transcriptions across different datasets directly, using string identity as a criterion to assess if two translations are identical or not. However, word form comparisons using phonetic data are still error-prone as datasets often differ with respect to details in the concrete realization of phonetic transcriptions (Anderson et al., 2018). Since variation in phonetic transcription is a norm rather than an exception (Anderson et al., 2023), we have to find a metric that allows us to distinguish those cases where two translations are identical even if the phonetic transcriptions dif-

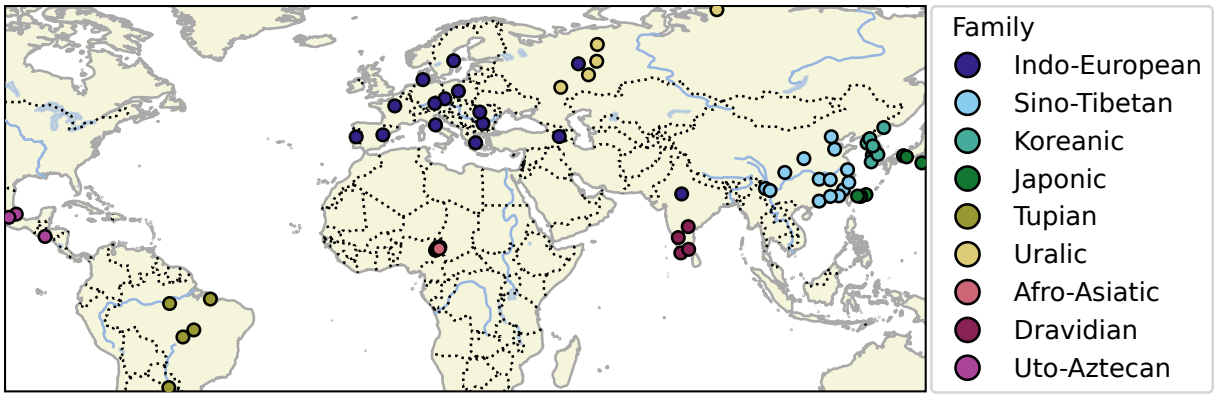


Figure 1: Location of the languages investigated in our study. For each of the 70 languages, two wordlists were identified in the Lexibank repository (for a full list of languages see file `values.csv` in the supplement).

fer slightly from those cases where two translations point to two different words. In order to address this problem, we decided for an automated approach that measures the phonetic similarity of sound sequences, rather than their identity with respect to the symbols used in transcriptions.

Our approach makes use of the Sound-Class Based Phonetic Alignments algorithm (SCA, List, 2012b), which was originally designed to align words phonetically, but which comes along with a measure for phonetic distances that ignores minor transcription differences. SCA distance scores derived from phonetic alignments carried out with the help of the SCA algorithm, have been shown to work quite well in the task of automated cognate detection (List, 2012a) and borrowing identification (Miller and List, 2023). SCA distance scores range between 0 (near identity of phonetic sequences) and 1 (very low similarity). Assuming that a score below 0.5 points to differences resulting from phonetic transcriptions, while scores higher than 0.5 result from differences stemming from different translations, we employ SCA distances as a proxy to detect whether two translations are *similar* (reflecting only phonetic variation) or *different* (reflecting true differences in translation).

In addition to the automated identification of translation differences with the help of SCA distances, we also computed the average SCA distances for all language pairs in our sample, as well as the edit distance (also known as *Levenshtein distance*, Levenshtein, 1966), both in its original and its normalized form (where we divide the distance by the longer of the two sequences, see List 2014, 178). All in all, these metrics should allow us to assess the amount of differences in concept translations across multilingual wordlists fairly well.

3.4 Preprocessing

Lexibank occasionally contains data in which morpheme boundaries are marked with the help of a plus symbol (+). Since morpheme boundaries would unnecessarily confuse phonetic alignment algorithms, introducing extra noise that we are not primarily interested in, we deleted all morpheme boundary markers from the sound sequences before comparing them. For the same reason, we also decided to ignore tone markers in the data. These do not occur in all datasets, but are instead mostly restricted to South-East Asian languages. Since tone annotation in phonetic transcription can vary considerably, probably even more than the transcription of consonants and vowels, we also ignored all tones in datasets from South-East Asian languages.

3.5 Evaluation

In order to test whether this approach is suitable to provide useful information on translation differences, we created a test set. Using EDICTOR (List et al., 2025b), translation differences were annotated for all language and concept pairs in the Indo-European datasets of our sample. This allowed us to use the manual annotations as a gold standard and to assess the suitability of SCA distance scores to identify true differences in concept translation. Results of this comparison are reported in the form of precision, recall, and F-Scores (List, 2014, 191–192).

3.6 Implementation

All methods are implemented in Python. We use LingPy to compute SCA distances and edit distances (https://pypi.org/project/lingpy, Version 2.6.13, List and Forkel, 2023). For the han-

dling of wordlist data in CLDF format, we use CL Toolkit (https://pypi.org/project/cltoolkit, Version 0.2.0, List and Forkel, 2022). For the inspection and manual annotation of wordlist data, EDICTOR was used (List et al., 2025b). For the creation of the map in Figure 1, CLDFViz was used (Version 1.3.0, Forkel, 2024). In order to download the data in its most recent versions defined by the Lexibank 2 repository (Blum et al., 2025a), we used PyLexibench (Häuser et al., 2025, https://pypi.org/project/pylexibench).

4 Results

4.1 Handling Phonetic Variation

As mentioned before, we are interested in finding a way to identify words that reflect the same original word form that is represented in slight variations in the phonetic transcriptions. Having a reliable automated approach for this task is important in order to allow us to investigate differences in concept translation on a larger amount of data.

To test how well the SCA distances can help us to identify words that only vary by phonetic transcription, we conducted a test study in which we annotated all hand-selected language pairs from the Indo-European group, indicating for each pair of words whether they were identical – despite potentially diverging phonetic transcriptions – or different. This dataset of 15 manually annotated language pairs was then compared with the results derived from our automated approach using the SCA distance. The results of this analysis are shown in Table 2. As the table illustrates, there are only minor differences between the automated and the manual annotation, with F-Scores of 0.98 on average for all 15 language pairs. We conclude from this that using SCA distances with a threshold of 0.5 provides a very good approximation to distinguish between identical word forms with potentially diverging transcriptions and word forms that reflect true differences in concept translation, allowing us to derive major conclusions when testing it on additional datasets.

Language	Precision	Recall	F-Score
Armenian (Eastern)	0.99	0.99	0.99
Bulgarian	1.00	0.99	0.99
Czech	1.00	0.99	1.00
Danish	1.00	0.95	0.97
French	0.99	0.99	0.99
German	1.00	0.95	0.98
Greek	0.97	0.99	0.98
Hindi	0.99	0.99	0.99
Polish	1.00	0.99	1.00
Portuguese	1.00	0.96	0.98
Romanian	0.98	0.96	0.97
Russian	1.00	0.99	0.99
Spanish	1.00	0.98	0.99
Swedish	0.99	0.97	0.98
Italian	0.97	0.96	0.96
TOTAL	0.99	0.98	0.98

Table 2: Comparison of the evaluation study on Indo-European. Precision and recall are calculated per concept, counting true and false negatives and positives for all possible pairings within the same concept slot for identical languages. F-Scores are based on the harmonic mean calculated from precision and recall.

4.2 Variation in Concept Translation

We conducted two tests on the 10 datasets. During the first test, we compared all languages by matching Glottocodes with each other. If more than one variety was assigned the same Glottocode in a given dataset, all possible pairs were assembled and average values for word pair identity, similarity, and SCA distances were computed. In the second experiment, only those language pairs were considered that we had identified as reflecting the same varieties (to the best of our knowledge). In all cases, we computed the identity of the sound sequences that appeared as translations for the same concepts, the *similarity* (with those pairs defined as similar whose SCA distance was below our threshold of 0.5), and the SCA distances. If more than one translation was available for the same concept in a given language variety, all possible pairs were compared and the average value of the individual scores was computed.

The results of the comparison based on matching Glottocodes are shown in Table 3, those of the comparison based on hand-selected language pairs are in Table 4. As can be seen from the tables, the results do not differ too much from each other, at least as far as their tendencies are concerned. Nevertheless, a closer inspection of the differences between the two approaches reveals that the linking of languages to Glottocodes shows some major problems in the datasets on the Chadic, Japonic, and Koreanic groups. In all three dataset pairs (Japonic and Koreanic data in one pair are both taken from the same study of Robbeets et al. 2021), we find that links to Glottolog could be improved. The data by Robbeets et al. (2021), for example, provides the

Family	Pairs	↑ Identical	STD	↑ Similar	STD	↓ SCA	STD	↓ NED	STD	↓ ED	STD
Bai	2	0.32	0.12	0.83	0.02	0.20	0.04	0.44	0.10	**1.33**	0.32
Chadic	10	0.07	0.05	0.71	0.14	0.32	0.10	0.54	0.11	3.34	0.63
Chinese	12	**0.39**	0.11	0.77	0.05	0.24	0.04	0.38	0.06	1.66	0.25
Dravidian	4	0.06	0.04	0.79	0.02	0.25	0.03	0.57	0.08	3.30	0.44
Indo-European	15	0.31	0.22	**0.94**	0.03	**0.09**	0.05	0.30	0.13	1.48	0.56
Japonic	7	0.28	0.19	0.74	0.24	0.27	0.17	0.40	0.20	2.09	1.02
Koreanic	9	0.06	0.04	0.74	0.21	0.31	0.14	0.59	0.12	3.01	0.72
Tupian	6	0.32	0.19	0.73	0.32	0.29	0.21	0.39	0.25	2.01	1.22
Uralic	6	0.19	0.12	0.85	0.06	0.17	0.08	0.42	0.13	2.26	0.78
Uto-Aztecan	4	0.34	0.22	0.88	0.03	0.18	0.06	**0.32**	0.11	1.79	0.85
TOTAL	75	0.23	0.13	0.80	0.07	0.23	0.07	0.44	0.10	2.23	0.74

Table 3: Major results per dataset for our comparative study, comparing all languages that show matching Glottocodes with each other, in terms of Identical (phonetic strings match perfectly) and Similar word pairs (phonetic strings show SCA distance beyond our threshold of 0.5), as well as averaged SCA distances, normalized edit distance, and traditional edit distances. Highest similarities are marked in bold font, lowest similarities are shaded in gray.

same Glottocodes for historical and modern varieties of Japanese and Korean. Any comparison that matches solely by Glottocode will therefore run the risk of comparing data from different stages of the same language. This example shows that scholars who base their analyses on Glottocodes should take particular care in selecting the most representative varieties. Especially when aggregating languages from different sources, one should make sure to provide extra checks on top of CLDF that would ensure that the same language varieties are being compared. Matching Glottocodes are an extremely good proxy, but they do not provide a guarantee that languages varieties from different sources really match.

What we can also see from the table is that we have extremely low values on phonetic identity in both comparisons, while phonetic similarity (as reflected in SCA distance scores beyond the value 0.5) shows a drastic increase. This proves the usefulness of computing SCA distance scores instead of comparing whether sound sequences are fully identical or not. It also shows (as we have already seen in the previous section) that SCA distances seem to provide quite sensitive results when it comes to assessing the near-identity of sound sequences reflecting slight transcription differences.

When considering only the phonetic similarity scores, which point to differences in lexeme choice when it comes to translating a concept into a given language variety, it seems remarkable that – apart from Indo-European, where differences make up only 6% – we find that differences between wordlists that we would expect to represent identical language varieties show a considerable amount of variation. On average, only 83% of all word pairs for our hand-selected sample of language pairs taken from different sources seem to be truly the same. In the remaining 17% of cases, we find that the concepts were translated differently. Recalling that Geisler and List (2010) report differences of about 10% with respect to lexeme choice in the Romance partition of the Indo-European dataset they compared, we can conclude that the numbers are even worse when looking at data from more language families.

That Indo-European data shows the lowest variation in our study should not come as a surprise. First, the language family has been studied in much more detail than any other language family in the world. Second, the principles by which Heggarty et al. (2023) compiled their data have been heavily influenced by the principles laid out by the Moscow school of historical linguistics (Kassian et al., 2010), from which the second dataset on Indo-European languages in our sample was taken (Starostin, 2005). Figure 2 illustrates the dominance of Indo-European, in representing all 70 pair comparisons in a bar chart with increasing lexical variation. As can be seen from the example, with the exception of the Indo-European datasets that show the highest similarity with respect to the translation of the concept lists into the target languages, there is no real trend that might hint to major problems with the data in particular language groups or language families. Instead, it seems that what we find in our experiments can be considered as the kind of variation that one would expect when considering the task of having independent people translate a concept list into the same language.

Family	Pairs	↑ Identical	STD	↑ Similar	STD	↓ SCA	STD	↓ NED	STD	↓ ED	STD
Bai	2	0.32	0.12	0.83	0.02	0.20	0.04	0.44	0.10	**1.33**	0.32
Chadic	5	0.11	0.03	0.79	0.04	0.27	0.02	0.48	0.06	3.00	0.36
Chinese	12	**0.39**	0.11	0.77	0.05	0.24	0.04	0.38	0.06	1.66	0.25
Dravidian	4	0.06	0.04	0.79	0.02	0.25	0.03	0.57	0.08	3.30	0.44
Indo-European	15	0.31	0.22	**0.94**	0.03	**0.08**	0.03	**0.29**	0.11	1.39	0.51
Japonic	6	0.31	0.18	0.83	0.05	0.21	0.07	0.34	0.10	1.75	0.54
Koreanic	13	0.06	0.03	0.82	0.09	0.25	0.06	0.54	0.06	2.76	0.45
Tupian	5	0.38	0.13	0.86	0.05	0.21	0.05	0.29	0.05	1.53	0.34
Uralic	5	0.23	0.08	0.87	0.02	0.14	0.04	0.37	0.06	2.00	0.50
Uto-Aztecan	3	0.24	0.14	0.87	0.01	0.20	0.04	0.37	0.08	2.16	0.54
TOTAL	70	0.24	0.12	0.84	0.05	0.21	0.06	0.41	0.10	2.09	0.70

Table 4: Major results per dataset for our comparative study, comparing all 70 hand-selected language pairs in terms of Identical (phonetic strings match perfectly) and Similar word pairs (phonetic strings show SCA distance beyond our threshold of 0.5), as well as averaged SCA distances, normalized edit distance, and traditional edit distances. Highest similarities are marked in bold font, lowest similarities are shaded in gray.

4.3 Types of Variation in Concept Translation

In order to gain a better understanding of the sources of variation in concept translation, we carried out a more detailed analysis of the differences in the Indo-European and the Tupian data. From this comparison, we can identify two major kinds of translation differences. First, translations can point to completely different words. Second, translations may reflect the same word, but they differ morphologically.

As an example for completely different words provided as translations for the same concept in the same target languages, consider cases like the concept MEAT, translated correctly into French as [vjɑ̃ːd] in the Indo-European dataset by Heggarty et al. (2023), while the Indo-European dataset by Starostin (2005) provides two translations, [vjɑ̃d] and [ʃɛʁ], the latter pointing specifically to "flesh". Since one of the two forms is identical with the form in the first dataset, we count this translation difference as 0.5 in our calculations. We can see that the reason for the additional translation in the dataset by Starostin (2005) results from a lack of specification in the concept that was being compared. As an additional example, consider Tupian data for Paraguayan Guaraní, where [ɡwasu] in Galucio et al. (2015) is glossed as "big" (cf. Estigarribia, 2020, 56) in the original data and matched with the Concepticon gloss BIG in the Lexibank dataset. This word form may also be rendered as "be big", as in Gregores and Suárez (1967, 224), if one does not recognize in Paraguayan Guaraní the existence of an adjective class, for which "little evidence exists" (Estigarribia, 2020, 15). Conversely, Ferraz Gerardi and Reichert (2021) provide the word form [posogue] for the concept BIG, which could be translated into English as "huge" or "gigantic" (Guasch and Ortiz, 2008, 720). This example illustrates how subtle semantic differences influence lexical choice, leading to differences in the lexical forms selected for inclusion in a wordlist. The use of periphrastic phrasing with multiple word forms in contrast to a single word form also contributes to differences in concept translation between the datasets. Take the concept STAND in Italian as an example. Heggarty et al. (2023) provide the form [staːreimpjɛːdi], which consists of three word forms: a verb, a preposition and a noun. By contrast, Starostin (2005) specifies the single verb form [stare] for the same concept. In particular, for systems which implement automatic cognate detection, such as SCA, this kind of concept translation dissimilarity can be problematic as they often rely on surface phonetic similarity between comparative word forms. Many true cognates could therefore be missed, even if a common word form is present in the forms being compared.

Morphological variation also causes notable differences in concept translation. In Tupian, this involves a lack of consensus on the citation form of inflected parts of speech (mainly nouns and verbs) across different languages, associated with a limited availability of complete inflectional paradigms. Generally, the two wordlists differ in their tendency to report inflected forms (Galucio et al., 2015) in contrast to roots (Ferraz Gerardi and Reichert, 2021). As an example, consider cases of different morphological forms in Paraguayan Guaraní. Here, some nouns have what are traditionally considered alternating roots with initial consonant [t], [h], or

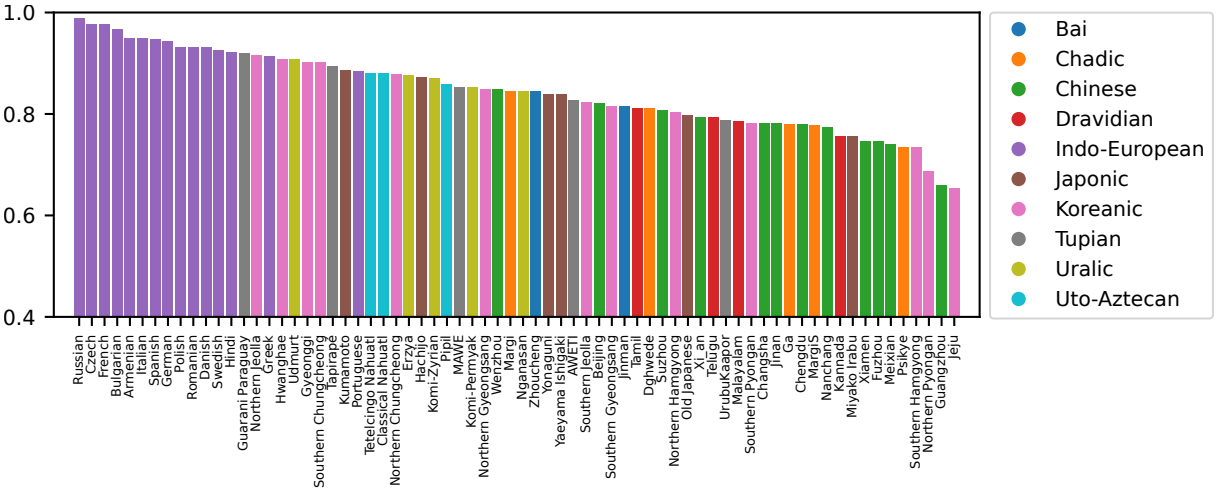


Figure 2: Comparison of all language pairs in the sample. The plot shows the hand-curated collection of 70 language pairs in our sample, colored by language family. The scores on the y-axis indicate the identified proportion of similar word pairs, showing an SCA distance beyond 0.5, ranked by their scores from left (high similarity) to right (low similarity).

[r]. The initial consonant may be treated as a prefix (Estigarribia, 2020, 63), with [t-] expressing, for instance, the "absolute or non-possessed form". Galucio et al. (2015) have a tendency to report the Paraguayan Guaraní forms with [t-], while Ferraz Gerardi and Reichert (2021) tend to only report the prefixless root. This issue is not unique to Tupian languages; similar mismatches arise from differing representations of inflectional morphology in other datasets. For example, in the Czech data for the concept SAY Heggarty et al. (2023) use the perfective form [r̥iːtst], whereas Starostin (2005) provides the imperfective form [riːkat]. Root allomorphy in Paraguayan Guaraní also contributes to differences between Galucio et al. (2015) and Ferraz Gerardi and Reichert (2021). Take, for example, the concept COME in Paraguayan Guaraní. It exhibits root allomorphy, appearing as [ju] or [u] (Estigarribia, 2020, 146–147). Galucio et al. (2015) represent both allomorphs whereas Ferraz Gerardi and Reichert (2021) only include [ju]. While a manual analysis would most likely treat these differences as marginal, characterizing the variants as identical translations of the original concept into the target language in question, there are cases where morphological variation can yield substantial differences in the translations. As an example, consider the two forms for YELLOW in Urubú-Ka'apór. Based on Kakumasu and Kakumasu (2007, 56, 116, 200, 204), the representation in Galucio et al. (2015) is a suffix meaning "yellow" [-ju], while Ferraz Gerardi and Reichert (2021) indicate the third person form of the "descriptive verb" "be yellow" [itawa].

5 Discussion and Conclusion

Our results confirm the findings of Geisler and List (2010), who emphasize that concept translation may cause larger discrepancies between wordlists compiled independently by different scholars. While Geisler and List (2010) report differences of about 10% for the Romance language partition in the two Indo-European datasets they compared, our results, taking data from 9 different language families into account, show that the differences are typically even larger, with only 83% of concept translations reflecting the same underlying word forms on average. Figure 2 also demonstrates the magnitude of variation which occurs in concept translation between individual language varieties, also within the same language family, with some varieties such as Russian achieving the exceptionally high similarity score of 0.99 and others like Jeju scoring 0.65. Many factors of course impact this variation, not least the increased standardization present in certain language varieties we examined, but also perhaps sociolinguistic factors such as dialect contact, semantic extension and cultural salience, which is beyond the scope of our current study. It is also worth noting that certain language families such as Tupian also contain many unknowns, it is therefore expected to have a certain level of disparity in these datasets. While we have not tested the impact that differences in concept

translation may have on phylogenetic reconstruction, we consider our results robust enough to call for the attention of scholars who use phylogenetic methods to answer big picture questions about human prehistory.

Although we think it is far too early to discard phylogenetic methods, we think that it would be useful for phylogenetic approaches to take potential problems resulting from the concept translation stage during the wordlist compilation seriously and make sure to apply certain measures to increase the robustness of their inferences. These measures could include tests on sampling errors, as outlined in Feld and Maxwell (2019), involve additional tests on inter-annotator agreement (McDonald et al., 2019) during concept translation, or entail conducting robustness tests similar to the bootstrap in phylogenetic reconstruction.

In any case, the declared goal of phylogenetic approaches should be the same as for traditional historical linguistics. Ratcliffe (2012) suggests that an ideal test of the reliability of the traditional comparative method for linguistic reconstruction would be to "take two teams of researchers trained in the comparative method, put them in the libraries, keep them in isolation from each other and see what they come up with" (Ratcliffe, 2012, 240). While it is clear that such experiments are still lacking in traditional historical linguistics, it seems important that scholars working in the field of historical language comparison, no matter if they work computationally or manually, maintain a mindset that does not take the reliability of their data for granted.

As far as our own experiments are concerned, we are still in the early stages of this research. Additional experiments – potentially even including additional data from language families that do not feature in the sample presented here – will be needed to get a better understanding of the potential impact of concept translation on phylogenetic analysis. It is possible that phylogenetic methods turn out to be robust enough to yield similar results in terms of subgrouping and divergence time estimates, even if the wordlists that they employed show a certain amount of differences in the translations. Without detailed analyses, however, we cannot be sure, and should not exclude the possibility that concept translation has a direct impact on the results of phylogenetic reconstruction analyses.

In addition to phylogenetic reconstruction, it would also be important to test to which degree concept translation might influence the results of other studies that make use of multilingual wordlists. Among these, the most important candidates that we can identify are global studies on sound symbolism or similar phenomena (Wichmann et al., 2010; Johansson et al., 2020), as well as studies that make use of cross-linguistic colexification data (Jackson et al., 2019; Tjuka et al., 2024; Rubehn and List, 2025). Recently introduced methods for language affiliation without phylogenetic reconstruction (Blum et al., 2025b) may also suffer from variation in concept translation. In any case, the last word on the robustness of multilingual wordlists has not yet been spoken, and more tests are needed to understand the full implications of the findings that we reported in this study.

Supplementary Material

The data, code, and detailed instructions needed to replicate this study are curated on Codeberg (https://codeberg.org/calc/concept-translation-study, Version 1.0) and archived with Zenodo (https://doi.org/10.5281/zenodo.15653036).

Limitations

The most important limitation to our current study is the uneven distribution of data across language families and varieties. For instance, we have 15 comparative varieties for Indo-European, while we only have two for Bai. There is also a disparity in the number of matching concepts between matching language varieties which allows us to achieve a more extensive sample of certain paired varieties. At the moment, we do not see how these limitations could be addressed consistently. In the long run, it seems that we must try to increase the number of comparisons by trying to identify more datasets that were independently compiled for the same languages.

Funding Information

This project was supported by the ERC Consolidator Grant ProduSemy (PI Johann-Mattis List, Grant No. 101044282, see https://doi.org/10.3030/101044282). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

Author Contributions

DS and JML initiated the study. DS coded the data, LC, AR, and KPvD provided assistance in questions on data. AR and KPvD provided assistance in questions on the code, DS and JML wrote the code base and analyzed the data. DS and JML wrote the first draft, LC contributed manual analyses of the results on Tupían and Indo-European languages and assisted in writing. All authors read and commented on the revised draft and agree with its final contents.

References

Bryan Allen. 2024. *CLDF dataset derived from Allen's "Bai dialect survey" from 2007*. Zenodo, Geneva.

Cormac Anderson, Tiago Tresoldi, Thiago Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting*, 4(1):21–53.

Cormac Anderson, Tiago Tresoldi, Simon J. Greenhill, Robert Forkel, Russell Gray, and Johann-Mattis List. 2023. Variation in phoneme inventories: Quantifying the problem and improving comparability. *Journal of Language Evolution*, 8(2):149–168.

Beijing University. 2024. *CLDF dataset derived from Beijing University's "Chinese dialect vocabularies" from 1964*. Zenodo, Geneva.

Frederic Blum, Carlos Barrientos, Johannes Englisch, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, and Johann-Mattis List. 2025a. Lexibank 2: Precomputed features for large-scale lexical data [version 1; peer review: 3 approved]. *Open Research Europe*, 5:126.

Frederic Blum, Steffen Herbold, and Johann-Mattis List. 2025b. From isolates to families: Using neural networks for automated language affiliation. In *Proceedings of the Association for Computational Linguistics 2025*, pages 1–12.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Will Chang, Chundra Cathcart, David Hall, and Andrew Garret. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244.

Bǎoyà Chén. 1996. *Lùn yǔyán jiēchù yǔ yǔyán liánméng [Language contact and language unions]*. Yǔwén, Běijīng.

Albert Davletshin. 2024. *CLDF dataset derived from Davletshin's "Proto-Aztecan languages" from 2012*. Zenodo, Geneva.

Johannes Dellert. 2024. *CLDF dataset derived from Dellert et al.'s "NorthEuraLex (version 0.9)" from 2020*. Zenodo, Geneva.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2020. NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*, 54:273–301.

Mark Donohue, Tim Denham, and Stephen Oppenheimer. 2012. New methodologies for historical linguistics? Calibrating a lexicon-based methodology for diffusion vs. subgrouping. *Diachronica*, 29(4):505–522.

Bruno Estigarribia. 2020. *A grammar of Paraguayan Guarani*. University College London Press, London.

Jan Feld and Alexander Maxwell. 2019. Sampling error in lexicostatistical measurements: A Slavic case study. *Diachronica*, 36(1):100–120.

Fabrício Ferraz Gerardi and Stanislav Reichert. 2021. The Tupí-Guaraní language family: A phylogenetic classification. *Diachronica*, 38(2):151–188.

Fabrício Ferraz Gerardi and Stanislav Reichert. 2024. *CLDF dataset derived from Gerardi and Reichert's "The Tupí-Guaraní language family: A phylogenetic classification" from 2021*. Zenodo, Geneva.

Robert Forkel. 2024. *CLDFViz: A Python library providing tools to visualize data from CLDF datasets [Software Library, Version 1.3.0]*. Zenodo, Geneva.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205.

Ana Vilacy Galucio, Sérgio Meira, Joshua Birchall, Denny Moore, Nilson Gabas Júnior, Sebastian Drude, Luciana Storto, Gessiane Picanço, and Carmen Reis Rodrigues. 2015. Genealogical relations and lexical distances within the Tupian linguistic family. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, 10(2):229–274.

Ana Vilacy Galucio, Sérgio Meira, Joshua Birchall, Denny Moore, Nilson Gabas Júnior, Sebastian Drude, Luciana Storto, Gessiane Picanço, and Carmen Reis Rodrigues. 2024. *CLDF dataset derived from Galucio et al.'s "Lexical distances within the Tupian linguistic family" from 2015*. Zenodo, Geneva.

Hans Geisler and Johann-Mattis List. 2010. Beautiful trees on unstable ground: Notes on the data problem in lexicostatistics. In Heinrich Hettrich, editor, *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik*. Reichert, Wiesbaden. The document was submitted in 2010 and is still waiting for publication.

Hans Geisler and Johann-Mattis List. 2022. Of word families and language trees: New and old metaphors in studies on language history. *Moderna*, 24(1-2):134–148.

Richard Gravina. 2024. *CLDF dataset derived from Gravinas's "Proto-Central Chadic" from 2014*. Zenodo, Geneva.

Simon J. Greenhill, Hannah J. Haynie, Robert M. Ross, Angela M. Chira, Johann-Mattis List, Lyle Campbell, Carlos A. Botero, and Russell D. Gray. 2025. *CLDF dataset accompanying Greenhill et al.'s "Origin of Uto-Aztecan" from 2022*. Zenodo, Geneva.

Emma Gregores and Jorge A Suárez. 1967. *A description of colloquial Guaraní*, volume 27 of *Janua Linguarum. Series Practica*. Mouton & Co., The Hague and Paris.

Antonio Guasch and Diego Ortiz. 2008. *Diccionario Castellano–Guaraní Guaraní–Castellano. Sintáctico-fraseológico-ideológico*, 13th edition. CEPAG, Asunción.

Harald Hammarström, Martin Haspelmath, Robert Forkel, and Sebastian Bank. 2024. *Glottolog [Dataset, Version 5.1]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Luise Häuser, Robert Forkel, and Johann-Mattis List. 2025. Pylexibench — Generating data for Lexibench with a Python package. *Computer-Assisted Language Comparison in Practice*, 8:25–37.

Luise Häuser, Gerhard Jäger, and Alexandros Stamatakis. 2024. Computational approaches for integrating out subjectivity in cognate synonym selection. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 162–172, Irvine, CA. Association for Computational Linguistics.

Paul Heggarty, Cormac Anderson, and Matthew Scarborough. 2024. *CLDF dataset derived from Heggarty et al.'s "Indo-European cognate relationships database" from 2023*. Zenodo, Geneva.

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechoslaw Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F. Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, Matthew Boutilier, Cassandra Freiberg, Robert Tegethoff, Matilde Serangeli, Nikos Liosis, Krzysztof Stroński, Kim Schulte, Ganesh Kumar Gupta, Wolfgang Haak, Johannes Krause, Quentin D. Atkinson, Simon J. Greenhill, Denise Kühnert, and Russell D. Gray. 2023. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656).

Hans J. Holm. 2007. The new arboretum of Indo-European "trees": Can new algorithms reveal the phylogeny and even prehistory of Indo-European? *Journal of Quantitative Linguistics*, 14(2-3):167–214.

Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Peter J. Mucha, Robert Forkel, Simon J. Greenhill, Russell D. Gray, and Kristen Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Niklas Erben Johansson, Andrey Anikin, Gerd Carling, and Arthur Holmer. 2020. The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguistic Typology*, 24(2):253–310.

James Y. Kakumasu and Kiyoko Kakumasu. 2007. *Dicionário por tópicos Kaapor-Português*. Associação Internacional de Lingüística—SIL Brasil, Cuiabá.

Alexei S. Kassian, George Starostin, Anna Dybo, and Vasiliy Chernov. 2010. The Swadesh wordlist: An attempt at semantic specification. *Journal of Language Relationships*, 4:46–89.

Alexei S. Kassian, Mikhail Zhivlov, George Starostin, Artem A. Trofimov, Petr A. Kocharov, Anna Kuritsyna, and Mikhail N. Saenko. 2021. Rapid radiation of the inner Indo-European languages: An advanced approach to Indo-European lexicostatistics. *Linguistics*, 59(4):949–979.

Vishnupriya Kolipakam. 2024. *CLDF dataset derived from Kolipakam et al.'s "DravLex:" from 2018*. Zenodo, Geneva.

Charles H. Kraft. 2024. *CLDF dataset derived from Kraft's "Chadic wordlists" from 1981*. Zenodo, Geneva.

Sean Lee. 2024. *CLDF dataset derived from Lee's "Sketch of language history in the Korean peninsula" from 2015*. Zenodo, Geneva.

Sean Lee and Toshikazu Hasegawa. 2024. *CLDF dataset derived from Lee and Hasegawa's "Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages" from 2011*. Zenodo, Geneva.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Johann-Mattis List. 2012a. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, pages 117–125, Stroudsburg.

Johann-Mattis List. 2012b. SCA: Phonetic alignment based on sound classes. In Marija Slavkovik and Dan Lassiter, editors, *New directions in logic, language and computation*, pages 32–51. Springer, Berlin and Heidelberg.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Johann-Mattis List. 2018. Tossing coins: Linguistic phylogenies and extensive synonymy. *The Genealogical World of Phylogenetic Networks*, 7(2).

Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2024. *Cross-Linguistic Transcription Systems [Dataset, Version 2.3.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List and Robert Forkel. 2022. *CL Toolkit: A Python library for the processing of cross-linguistic data [Software Library, Version 0.2.0]*. Zenodo, Geneva.

Johann-Mattis List and Robert Forkel. 2023. *LingPy: A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.13]*. MCL Chair at the University of Passau, Passau.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(316):1–31.

Johann-Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon J. Greenhill, and Robert Forkel, editors. 2025a. *CLLD concepticon 3.3.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Kellen Parker van Dam, and Frederic Blum. 2025b. *EDICTOR 3: An interactive tool for computer-assisted language comparison [Software Tool, Version 3.1]*. MCL Chair at the University of Passau, Passau.

Liú Lìlǐ, Wáng Hóngzhōng, and Bǎi Yíng. 2024. *CLDF dataset derived from Liú et al.'s "Collection of basic words in Chinese dialects" from 2007*. Zenodo, Geneva.

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23.

April McMahon, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005. Swadesh sublists and the benefits of borrowing: An Andean case study. *Transactions of the Philological Society*, 103:147–170.

John E. Miller and Johann-Mattis List. 2023. Detecting lexical borrowings from dominant languages in multilingual wordlists. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2599–2605, Dubrovnik, Croatia. Association for Computational Linguistics.

Robert R. Ratcliffe. 2012. On calculating the reliability of the comparative method at long and medium distances: Afroasiatic comparative lexica as a test case. *Journal of Historical Linguistics*, 2(2):239–281.

Martine Robbeets. 2025. *CLDF dataset derived from Robbeets et al.'s "Triangulation of Transeurasian languages" from 2021*. Zenodo, Geneva.

Martine Robbeets, Remco Bouckaert, Matthew Conte, Alexander Savelyev, Tao Li, Deog-Im An, Ken ichi Shinoda, Yinqiu Cui, Takamune Kawashima, Geonyoung Kim, Junzo Uchiyama, Joanna Dolińska, Sofia Oskolskaya, Ken-Yōjiro Yamano, Noriko Seguchi, Hirotaka Tomita, Hiroto Takamiya, Hideaki Kanzawa-Kiriyama, Hiroki Oota, Hajime Ishida, Ryosuke Kimura, Takehiro Sato, Jae-Hyun Kim, Bingcong Deng, Rasmus Bjørn, Seongha Rhee, Kyou-Dong Ahn, Ilya Gruntov, Olga Mazo, John R. Bentley, Ricardo Fernandes, Patrick Roberts, Ilona R. Bausch, Linda Gilaizeau, Minoru Yoneda, Mitsugu Kugai, Raffaela A. Bianco, Fan Zhang, Marie Himmel, Mark J. Hudson, and Chao Ning. 2021. Triangulation supports agricultural spread of the Transeurasian languages. *Nature*, 599(7886):616–621.

Arne Rubehn and Johann-Mattis List. 2025. Partial colexifications improve concept embeddings. In *Proceedings of the Association for Computational Linguistics 2025*, pages 1–15.

Arne Rubehn, Jessica Nieder, Robert Forkel, and Johann-Mattis List. 2024. Generating feature vectors from phonetic transcriptions in cross-linguistic data formats. *Proceedings of the Society for Computation in Linguistics*, 7(1):205–216.

Sergey A. Starostin. 2005. Indo-European files in DBF/VAR. In George Starostin, editor, *The Tower of Babel*. RGGU, Moscow.

Sergey. A. Starostin. 2024. *CLDF dataset derived from Starostin's "Indo-European files in DBV/VAR" from 2005*. Zenodo, Geneva.

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16:157–167.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, 96(4):452–463.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21:121–137.

Kaj Syrjänen, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski, and Niklas Wahlber. 2024. *CLDF dataset derived from Syrjänen et al.'s "Shedding more light on language classification" from 2013*. Zenodo, Geneva.

Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2024. Universal and cultural factors shape body part vocabularies. *Scientific Reports*, 14(10486):1–12.

Feng Wang. 2024. *CLDF dataset derived from Wang's "Language contact and language comparison" from 2004*. Zenodo, Geneva.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2010. Sound symbolism in basic vocabulary. *Entropy*, 12(4):844–858.

Annotating and Inferring Compositional Structures in Numeral Systems Across Languages

Arne Rubehn[1], Christoph Rzymski[2], Luca Ciucci[1], Katja Bocklage[1], Alžběta Kučerová[1], David Snee[1], Abishek Stephen[3], Kellen Parker van Dam[1], Johann-Mattis List[1]

[1]Chair for Multilingual Computational Linguistics, University of Passau, Passau, Germany
[2]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[3]Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic

Abstract

Numeral systems across the world's languages vary in fascinating ways, both regarding their synchronic structure and the diachronic processes that determined how they evolved in their current shape. For a proper comparison of numeral systems across different languages, however, it is important to code them in a standardized form that allows for the comparison of basic properties. Here, we present a simple but effective coding scheme for numeral annotation, along with a workflow that helps to code numeral systems in a computer-assisted manner, providing sample data for numerals from 1 to 40 in 25 typologically diverse languages. We perform a thorough analysis of the sample, focusing on the systematic comparison between the underlying and the surface morphological structure. We further experiment with automated models for morpheme segmentation, where we find allomorphy as the major reason for segmentation errors. Finally, we show that subword tokenization algorithms are not viable for discovering morphemes in low-resource scenarios.

1 Introduction

Numeral systems represented by the words for cardinal numbers used in counting are an interesting kind of linguistic data: they code a part of the lexicon of human languages that is potentially large and often exhibits a regularity that increases with higher numbers. Regularity is reflected in the *recycling* of linguistic material used to create higher numbers, where morphemes for smaller number words are often reused to motivate the formation of larger numerals. In addition, numeral systems are also maximally *distinctive*. Being used to distinguish ordinal numbers, we rarely find cases in which two distinct numbers are expressed by the same word form, even if numeral words themselves can have multiple meanings outside of the number domain (as can be easily seen when browsing number words in the *Database of Cross-Linguistic Colexifications*, Rzymski et al., 2020).

Another important aspect of numeral systems is that they are not created in an ad-hoc fashion but have instead often evolved over hundreds of years. The evolution can leave traces in numeral systems that counter-act former regularity, leading to allophonic variation in the morphemes that compose numeral words. Language contact can also feature as an important aspect of evolution, resulting in extreme cases where languages use two or more numeral systems in combination, reflecting different stages of their history.

The fact that most numeral systems are *compositional*, while at the same time being distinctive and discrete in their denotation, makes them an interesting test object for linguistic analyses that deal with lexical compositionality in the context of language change. While one would otherwise have to cope with problems resulting from various kinds of morphological and semantic variation, numeral systems can be seen as an ideal test ground for the annotation and inference of compositional structures in the lexicon of human languages. In the following, we will try to illustrate this point in more detail. After a short overview on numeral systems in the context of descriptive and computational linguistics (§ 2), we present a small collection of numeral systems along with methods that can be used to annotate numeral systems manually or to segment numeral words automatically into morphemes (§ 3). After testing these methods and reporting the results on our small cross-linguistic sample of numeral systems (§ 4) we discuss our findings and point to ideas for future work (§ 5).

2 Background

The cross-linguistic diversity of numeral systems has attracted the interest of scholars since Hervás y Panduro's comparative work (1786), which pre-

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 29–42
August 1, 2025 ©2025 Association for Computational Linguistics

sented data from missionaries on many then little-known languages. Today, the most comprehensive database on numerals is Chan (2024), who collected data on more than 5,000 varieties, often provided by linguists with first-hand experience of the respective languages. The constant increase in data has allowed for the study of numeral systems from a formal (see e.g. Brandt Corstius, 1968; Hurford, 1975) and a typological perspective. The latter approach reached a turning point with Greenberg's (1978) 54 generalizations, most of which stood the test of time (Comrie, 2020).

Even though their synchronic structure may be opaque, numeral systems are diachronically motivated and are built through a limited number of cross-linguistic strategies (Heine, 1997, 18-34). They typically combine a small set of morphemes (mainly numbers, but also linking elements) according to three parameters, including (1) the choice of the base(s), (2) the operations applied to the base(s), following the implicational hierarchy: addition < multiplication < subtraction / division; and (3) the order of the morphemes (Greenberg, 1978; Moravcsik, 2017, 459-461). Despite the presumed regularity and compositionality of numeral systems, they may occasionally display gaps and ambiguities (Comrie, 1997, 2005, 79-80).

The most common bases are 'five', 'ten', and 'twenty', whose conceptual sources are, respectively, the fingers of the hand, of both hands, and of all hands and toes (Heine, 1997, 19-24; on finger counting and its cultural variability, see Bender and Beller, 2012). Decimal systems are the most frequent worldwide, followed by vigesimal and quinary systems (Skirgård et al., 2023). Languages can employ more than one base, resulting in hybrid numeral systems.

While languages with no numerals or only the number 'one' are rare (Hammarström, 2010), the numeral systems of many languages, particularly in South America, New Guinea and Australia, are restricted to a few numerals (Moravcsik, 2017, 459). According to Dixon (2012, 71-72), this indicates that the speakers did not count and enumeration was not the primary use of these number words. Hammarström (2008) observed that pidgins and creoles tend to have more complex numeral systems than the global average, with their frequent origin as trade languages being a potential contributing factor. Numeral systems often developed out of contact, which usually comes with societal change, and borrowing may also involve the lowest numbers (Dixon, 2012, 75-77).

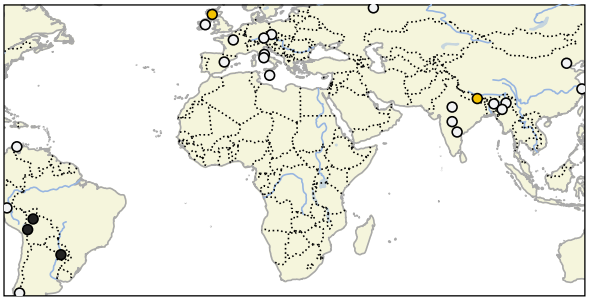


Figure 1: Geographical distribution of the languages in our sample, indicating the numeral bases they employ (white: 10, black: 5 and 10, orange: 10 and 20).

While numeral systems all over the world have been quite intensively investigated in the past, very few computational studies (Calude and Verkerk, 2016; Cathcart, 2025) formalize approaches to model compositionality and reveal motivation patterns underlying individual number words. Recent advances in the annotation of lexical motivation patterns (Hill and List, 2017) and the automated segmentation of words into morphemes (Goldsmith et al., 2017) open new possibilities for a computational investigation of numeral systems that we will discuss in more detail in the following.

3 Materials and Methods

3.1 Sample of Numeral Systems

We collected the cardinal numbers from 1 to 40 in 25 typologically diverse languages from Eurasia and Southern America, spanning ten different language families. Most language families, with the exception of Indo-European (12 languages) and Sino-Tibetan (5 languages), are represented by a single language. Table 1 provides a comprehensive overview of the languages covered in the sample, accompanied by a geographical visualization in Figure 1.

Most languages employ a decimal system, reflecting that the number 10 is by far the most common base. Three languages in our sample – Aymara, Cavineña, and Paraguayan Guaraní – make use of the number 5 as a base. They represent a hybrid between quinal and decimal systems, since the word for 10 is monomorphemic and used to express multiples of 10. Furthermore, two languages of our sample (Lamjung Yolmo and Scottish Gaelic) have retained a vigesimal system used in parallel to a

Family	Branch	Language	Base
Afro-Asiatic	Semitic	Maltese	10
Araucanian	—	Mapudungun	10
Arawak	Ta-Arawak	Wayuu	10
Aymaran	—	Aymara	5 / 10
Dravidian	Southern	Telugu	10
Indo-European	Balto-Slavic	Czech	10
		Russian	10
	Celtic	Irish	10
		Scottish Gaelic	10 / 20
	Germanic	German	10
	Indo-Iranian	Assamese	10
		Hindi	10
		Sanskrit	10
	Romance	French	10
		Italian	10
		Latin	10
		Spanish	10
Pano-Takanan	Takanan	Cavineña	5 / 10
Quechuan	Quechua I	Huallaga Quechua	10
Sino-Tibetan	Bodic	Lamjung Yolmo	10 / 20
	Brahmaputran	Uipo (Maringic)	10
	Patkaian	Makyam	10
	Sinitic	Mandarin Chinese	10
		Shanghainese	10
Tupian	Tupí-Guaraní	Paraguayan Guaraní	5 / 10

Table 1: Overview of languages covered in the sample, with their genetic classification and primary bases for counting.

decimal system, which results in alternating forms for numbers higher than 20.

All data were collected, annotated, and curated in a collaborative manner, such that the data for each language were thoroughly reviewed by at least two scholars: the responsible annotator for the given language, and at least one reviewer. The data were then aggregated and deployed as a unified dataset conforming to the Cross-Linguistic Data Formats (CLDF, Forkel et al., 2018; Forkel and List, 2020). Automated tests accounted for the structural integrity of the data (e.g. ensuring that one cognate ID does not map to more than one underlying form; the annotation format is described in detail in § 3.3).

3.2 Representing Numeral Systems in Tables

The CLDF specification builds on CSVW, a standard for tabular data on the web (https://csvw.org; Gower, 2021) that extends simple tabular data, typically represented in the form of CSV files, by metadata that can be used to specify the content of tabular data in various ways, including the combination of multiple tables in a relational database. Given that numeral data can be easily treated as *lexical data*, typically provided in the form of wordlists, we represent number systems as extended CLDF wordlists that build on the extended wordlist formats introduced by the Lexibank repository (List et al., 2022; Blum et al., 2025). Lexibank wordlists represent individual word forms as triples consisting of a *language*, a *concept*, and a *form*. In order to compare data from different sources, Lexibank makes use of reference catalogs that link language varieties to Glottolog (https://glottolog.org; Hammarström et al., 2025), map concepts to Concepticon (https://concepticon.clld.org; List et al., 2025a), and represent phonetic transcriptions compatible with the subset of the IPA proposed by the Cross-Linguistic Transcription Systems (CLTS) reference catalog (https://clts.clld.org; List et al., 2021).

While following Lexibank in assembling our exploratory database of numeral systems, we extend the format by adding new layers of annotation that help us to make individual analyses of the numeral systems explicit through annotation. As a first step, we rigorously split words into morphemes by adding morpheme boundary markers to all multi-morphemic words (using the plus symbol – + – as a boundary marker). As a second step, we identify language-internal partial cognates in all numeral systems in order to mark the degree by which morphemes are reused to build new numeral expressions (see List et al., 2016 on partial cognates). In other words, we annotate which morphemes recur across several forms by assigning a unique numerical ID to each morpheme. As a third step of analysis, we add *morpheme glosses* to the data to add human-readable semantic descriptors to all morphemes (Hill and List, 2017; Schweikhard and List, 2020). As a fourth step, we make use of *inline-alignments* in order to handle allomorphs by distinguishing underlying forms from surface forms (List, 2025). As a fifth step, we conduct *phonetic alignment analyses* (List, 2014) of all language-internal cognate morphemes, in order to facilitate the comparison of allomorphic variants that differ in length.

Table 2 shows how our annotations are rendered in tabular form, with examples for annotated numerals from German and French. The column *Segments* provides phonetic transcriptions, segmented into sounds, using a space as boundary marker, and secondarily segmented into morphemes, using the plus symbol as a boundary marker. The transcriptions use *inline alignments* (List, 2025) to align the surface forms with their underlying forms. Inline alignments (first introduced by List 2021 and later tested on Old Chinese etymologies by Pulini and

Language	Concept	Form	Segments	Cognates	Morphemes
German	one	*eins*	aɪ n s	1	ONE
German	two	*zwei*	ts v aɪ	2	TWO
German	three	*drei*	d r aɪ	3	THREE
German	twenty one	*einundzwanzig*	ai n -/s + ʊ n -/d + ts v a n + ts ɪ ç	1 4 5 6	ONE and TWEN TY
German	thirty two	*zweiunddreißig*	ts v aɪ + ʊ n -/d + d r aɪ + s/ts ɪ ç	2 4 3 6	THREE and THREE TY
French	one	*un*	œ̃	1	ONE
French	two	*deux*	d ø	2	TWO
French	three	*trois*	t ʁ w a	3	THREE
French	twenty one	*vingt-et-un*	v ɛ̃ t + e + œ̃	4 5 1	TWENTY and ONE
French	thirty two	*trente-deux*	t ʁ -/w -/ɑ + ɑ̃ t + d ø	3 6 2	THREE TY TWO

Table 2: Illustration of the format used to annotate morpheme boundaries along with allomorphic variation, language internal cognates, and morpheme glosses.

List 2024) use the slash symbol (/) in order to distinguish a surface sound (shown to the left of the slash) from its corresponding underlying sound. As an example, consider the transcription of German [aɪ n -/s] 'one' in the word for 'twenty one' in the table, where [s] is treated as the underlying form, while the surface form does not show this sound (which is marked by using the gap-symbol - before the slash). The notion of surface form and underlying form is strictly *technical*. We assume that one morpheme with multiple allomorphs has only one underlying form, which must consistently be aligned with all surface forms. We do not claim that this handling shows any cognitive or historical truth, but we aim for an annotation that would ideally be meaningful from a diachronic and cognitive perspective.

The columns *Cognates* and *Morphemes* provide information on language-internal cognates in the form of morphemes that are reused. Here, the *Cognates* column employs numerical identifiers, following the format proposed by List et al. (2016), while the functionally identical *Morphemes* column provides semantic glosses that help in making the lexical motivation underlying the formation of numerals transparent. This annotation, which provides explicit glosses for all morphemes constituting a word, was originally developed to make language-internal cognate relations more explicit (Hill and List, 2017). By now, however, it has been shown to be also very useful to provide rudimentary annotations of lexical motivation patterns (Brid et al., 2022).

3.3 Computer-Assisted Annotation

While the annotations shown in Table 2 can be easily carried out with the help of a spreadsheet editor or directly in text files, we use the web-based EDICTOR tool for the annotation of numeral data (List et al., 2025b). Originally, EDICTOR was designed to facilitate the process of creating multilingual comparative wordlists (List, 2017). Since Version 3.0 (List and van Dam, 2024), however, EDICTOR has been substantially extended to help with the annotation of lexical motivation patterns. Improvements include – among others – a visual rendering of inline alignments, sound sequences, cognate sets, and morpheme glosses, combined with annotation helpers for manual morpheme segmentation, as well as several sanity checks that increase the consistency of human annotation.

3.4 Automated Morpheme Segmentation

The task of unsupervised morpheme segmentation – automatically inferring a language's morphological structure from unannotated corpus data – has received notable attention in the field of Natural Language Processing, especially in the late 1990's and early 2000's (Hammarström and Borin, 2011). While those models were developed with a different background in mind, assuming the presence of relatively large training corpora, numeral systems naturally lend themselves as an interesting use case for morpheme segmentation models due to their high degree of compositionality. Therefore, we experiment with simple morpheme segmentation techniques to observe their performance in a transfer setting with much less data, but an extraordinarily strong morphological signal.

The first formalization of an algorithm for morpheme segmentation reaches back to Harris (1955) who proposed the so-called *Letter Successor Variety* (LSV) as a predictability measure at each position within a word. The underlying assumption is that the continuation of a word should be fairly predictable within a morpheme, but much harder to

predict at a morpheme boundary. Several proposals have been made to improve upon LSV. Hafer and Weiss (1974) suggest measuring predictability in terms of entropy rather than type variety (Letter Successor Entropy, LSE). They also propose Letter Predecessor Variety (or Entropy) as a logical inversion of LSV, processing each word backwards. Hammarström (2009) proposes *Letter Successor Max-Drop*, measuring how likely the most frequent continuation of a word is in comparison to all other potential continuations. Çöltekin (2010) suggests normalizing LSV by word position to account for the fact that LSV usually becomes smaller towards the end of a word. We experiment with all these different flavors of LSV, but report only LSE, since it performs best on average and all LSV variations show similar patterns in general. Following Hafer and Weiss (1974), we also experiment with a simple model that considers every possible prefix and suffix (in a computational sense) of a word form as a morpheme if and only if it appears as a complete word in the data. Using this simple measure, List (2023) reports promising results in inferring partial colexifications from multilingual wordlists which seem to advance concept embeddings substantially (Rubehn and List, 2025b).

A line of research that can be seen as complementary to LSV-based approaches formalizes the task of morpheme segmentation as a *minimum description length* (MDL) problem (Goldsmith, 2001). The basic idea behind MDL is to define a description length as a combination of basic tokens and rules to derive complex forms from the basic vocabulary. This notion is especially interesting on theoretical grounds, since the complexity of numeral systems can also be measured in terms of MDL (Hammarström, 2008). In an ideal setting, an MDL-based segmentation model is therefore expected to accurately infer and model the compositional structure of numeral systems. Representing this family of morpheme segmentation algorithms, we run our experiments with the Morfessor Baseline model (Creutz and Lagus, 2002, 2005; Virpioja et al., 2013).

3.5 Subword Tokenization

Algorithms for *subword tokenization* form an integral preprocessing step of state-of-the-art language models, since they effectively reduce the vocabulary size and avoid the occurrence of out-of-vocabulary items. While these tokenization methods in principle make downstream applications more flexible, it can at least be doubted whether the inferred subwords concord with the language's morphological structure (Batsuren et al., 2024). We apply three popular algorithms for subword tokenization on our multilingual numeral data: Byte-Pair-Encoding (BPE; Gage, 1994; Sennrich et al., 2016), WordPiece (Schuster and Nakajima, 2012), and Unigram tokenization (Kudo, 2018).

3.6 Evaluation

All models described in § 3.4 and § 3.5 are trained on unannotated and unsegmented representations of the numeral lists. The predicted segmentations are then evaluated against our manual annotations which serve as a gold standard. Since all models are inherently monolingual, each language is processed and evaluated independently.

Predicted segmentations can directly be evaluated against the gold standard using *precision* and *recall* (Virpioja et al., 2011). While we are aware of more sophisticated evaluation metrics for morphological analyses (Spiegler and Monson, 2010), we argue that simply calculating boundary precision and recall (BPR) is sufficient in our use case, since we investigate small corpora with hardly ambiguous morphological patterns. Due to its simplicity, BPR is readily interpretable, rendering it the ideal evaluation metric for our use case.

We run all experiments on two different representations of the numeral lists, relying on the *surface* and *underlying* forms respectively (see § 3.3 for details on the two representations). The former is a faithful representation of the actually observable word forms and therefore reflects a "real-world" use case for segmentation models. The latter is an artificially construed "ideal" setting that removes allophonic and allomorphic variation, that is, variation that needs to be explained on a different level than morphology. Comparing these two settings allows for a fine-grained evaluation of morpheme segmentation models, enabling us to assess the share of segmentation errors caused by allomorphy.

3.7 Implementation

The data were annotated using EDICTOR 3.1 (List et al., 2025b), and validated and compiled using CLDFBench (Forkel and List, 2020). The visualization in Figure 1 was created using CLDFViz (Forkel, 2024). All experiments regarding automated morpheme segmentation were run in Python, using LinSe (Forkel and List, 2024) to conveniently

	Average		Highest		Lowest	
	S	**U**	**S**	**U**	**S**	**U**
Morphemes	21.8	13.5	48	20	10	7
Expressivity	5.6	7.9	10.6	15	1.4	3.4
Opacity	1.60		3.18		1	
Code Length	2.53		3.83		1.68	

Table 3: Overview of statistics about the different numeral systems. **S** and **U** refer – where applicable – to surface vs. underlying forms.

represent the internal structure of word forms in different granularities. Morfessor was run from its Python package (Virpioja et al., 2013), all other models were implemented from scratch and are available through MorSeg, a package for morpheme segmentation in multi- and monolingual wordlists (Rubehn and List, 2025a). All data and code accompanying this study are made available in the supplementary material.

4 Analysis and Results

4.1 Sample Data of Coded Numeral Systems

Table 3 summarizes the results of computing different types of metrics based on surface and underlying forms across all languages in our sample (Table 6 in the appendix provides metrics for individual languages). In the table, we introduce three simple metrics – *expressivity*, *opacity*, and *length* – to get a better understanding of the data and the strategies to form higher numbers from basic morphemes. First, we measure the average *morpheme expressivity* of a language by counting how many different numbers are formed using this morpheme. For instance, Mandarin *wǔ* 'five' is used in the formation of the numbers 5, 15, 25, and 35 and therefore has an expressivity of 4. Expressivity is averaged over all morphemes found in a language's numeral system. For the rare cases where a language has multiple forms for the same number, expressivity is weighted accordingly. *Opacity* describes the ratio between allomorphic variants and morphemes, measuring the degree of allomorphy in a system. The lowest score is 1, with each morpheme in a language surfacing with the same form. Finally, the *average coding length* measures how many morphemes are used to form a word.

On theoretical grounds, the minimum amount of morphemes required in a numeral system is the base of that system. That means, a decimal system needs at least 10 different morphemes to be fully expressive. Indeed, our sample covers three languages – Mandarin, Mapudungun, and Huallaga Quechua – that use such a minimal decimal system to express the numbers up to 40. This observation holds true on both the surface and the underlying level, indicating that exactly these languages lack any kind of allomorphy. Mandarin is often taken as a prime example for a perfectly transparent and symmetric numeral system: Complex numerals are simply formed by concatenating the simple numbers from 1 to 10. For example, *twenty three* in Mandarin is *èr shí sān*, literally *two ten three* ($2 * 10 + 3$).

On the other side of the spectrum, we find Assamese with 20 different morphemes and Hindi with 48 distinct morphs, the highest value for the respective category. This aligns with the general impression that Indo-Aryan languages feature some of the most complex and opaque numeral systems of the world (Hammarström, 2008; Cathcart, 2025).

We observe a wide range of morpheme opacity. With Uipo, Huallaga Quechua, Mandarin, and Mapudungun, four languages in our sample have the lowest possible opacity of 1.0, thus not featuring any allomorphy in their numeral systems. On the other hand, the language with the highest opacity is still Hindi with a value of 2.82, followed by Lamjung Yolmo, Telugu, and Sanskrit. From these extreme cases, the impression might arise that the opacity correlates with the size of the underlying morpheme inventory. However, across the entire dataset, no significant correlation between these two metrics could be found.

The expressivity of morphemes and their allomorphic variants, on the other hand, shows a significant negative correlation with the number of morphemes. The interpretation is straightforward: The fewer morphemes are available in a system, the more expressive they need to be, and the more they will be used. It is therefore not surprising that exactly those three languages that employ a base of 5 (Aymara, Cavineña, and Paraguayan Guaraní) rank the highest in terms of expressivity on the surface and the underlying level. On the low end of expressivity, we again find Hindi and Assamese, as well as the modern Romance languages French, Italian, and Spanish.

Based on these correlations, one might expect that the average coding length is also directly dependent on the size of the morpheme inventory, since less available morphemes should – in theory – require longer word forms. However, no significant correlation between these two metrics could

be found. There is only a significant correlation between the coding length and the morpheme expressivity. Considering that our sample is heavily biased towards decimal systems, and that even the systems that employ other bases show traces of decimal coding, we cannot interpret these effects as a result of different numeral bases. Instead, this seems to result from oblique marking (connecting morphemes with particles like 'and' or 'with') which can happen independently of the numeral base.

Finally, we experiment with *type-token ratio* (TTE) and *entropy*, which have been proposed as measures of morphological complexity in the past (Bentz et al., 2017; Çöltekin and Rama, 2023). These metrics are not able to capture any aspect of complexity in our sample, since they correlate almost perfectly with the number of morphemes. We therefore conclude that in this special setting, TTE and entropy are dependent on the vocabulary size alone, which is probably due to the fact that morphemes in numeral systems by and large do not follow a Zipfian distribution, as is the case for words in natural language corpora.

4.2 Automated Morpheme Segmentation

Table 4 reports the overall performance of three models for automated morpheme segmentation on the individual languages, both for those cases where surface forms were passed to the algorithms, and where underlying forms were taken as the basis of analysis. From the model family based on Letter Successor Variety, we only report Letter Successor/Predecessor Entropy, which generally performed best.

The most obvious (and unsurprising) observation is that all models perform better on the underlying form than on the surface forms. Since it is a well-known issue in the literature that automated methods are challenged by allomorphy (Hammarström and Borin, 2011; Virpioja et al., 2011), this does not seem too surprising to us. Comparing the average scores of the models, however, shows that allomorphy is the biggest source of error for the analysis on surface forms, which naturally is the common use case for those models. By extension, it does not come as a surprise that opacity significantly correlates with how well the models perform on the surface forms, as shown in Figure 2.

Model	Surface Forms	Underlying Forms
Morfessor	0.74	0.88
LSPE	0.72	0.83
Affix	0.72	0.88

Table 4: Average F_1 scores of morpheme segmentation algorithms.

But even on the underlying forms – an "ideal" scenario in which allomorphy does not exist – there are notable differences in how well the morphological structure is detected by the models. Particularly interesting is the case of Uipo. This numeral system poses a big challenge for Morfessor and the Affix model, which both only achieve an F_1-score of 0.4 (while achieving a perfect precision of 1.0!). A closer look at the language data reveals that Uipo has a complex numeral system, in which even the numbers between 2 and 9 consist of two morphemes, a prefix and a stem. The number 6 for example is [t^h ɘ + r u k], but both morphemes are only used to form the number six (and by extension, numbers that are formed using 'six'). Without any further knowledge of the language, it is very hard if not impossible to recognize the underlying compositionality, leading to a massive undersegmentation by the models at hand. On the other hand, the high score of LSPE on Uipo – which may come as a surprise – can be described as a coincidental byproduct of the present morphophonology. As generally typical for South-East Asian languages, Uipo only allows the simple syllable structure CV(C), and each syllable in Uipo is a morpheme at the same time. Since there are more consonants than vowels, the continuation of a word is much less predictable at the start of a new syllable. LSPE can therefore accurately predict *syllable* boundaries, which happen to be morpheme boundaries as well.

On the other side, Morfessor is able to perfectly predict all morpheme boundaries in four languages at the surface level (Shanghainese, Mandarin, Huallaga Quechua, Mapudungun), and in seven more languages at the underlying level. Mapudungun seems to have a particularly transparent structure, since it is the only language that all three models segment perfectly at both representation levels. This makes Morfessor the model with the highest number of completely correct segmentations at the language level, showing that it clearly has the edge over the other two approaches tested, which is also indicated by the average performance. But even in this "ideal" scenario – no allomorphy and a system that shows clear compositional structures – Morfessor cannot accurately predict all morpheme boundaries for 14 out of 25 languages. For example, in

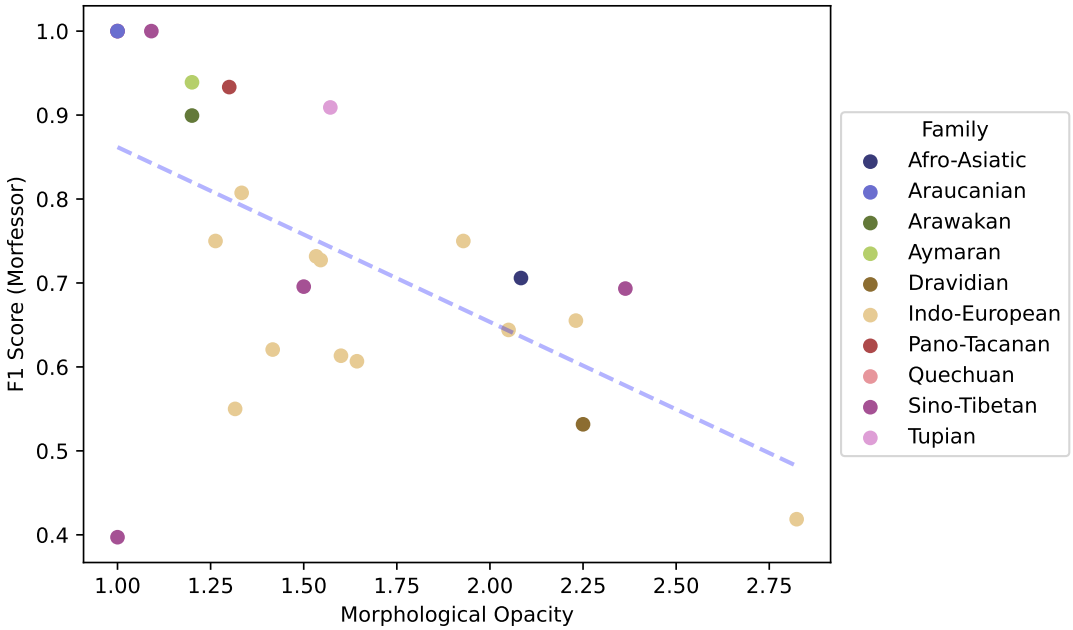


Figure 2: F1 scores of Morfessor on surface forms per language in correlation with the morphological opacity (Spearman's $\rho = -0.596$, p-value < 0.01).

the German words *zwan-zig* 'twen-ty' and *drei-ßig* 'thir-ty', the model fails to detect the morpheme boundaries, even in the underlying form where *-zig* [ts ɪ ç] and *-ßig* [s ɪ ç] are represented in the same way ({ts ɪ ç}). Generally, the model is much more prone to undersplitting than to oversplitting: On the underlying representation, it achieves a nearly perfect precision of 0.998, but a recall of only 0.80.

4.3 Subword Tokenization

Table 5 provides an overview of how accurately algorithms for subword tokenization can capture the morphological structure of the numeral systems at hand. It is evident that these models are in no way competitive with algorithms designed for the task of morphological segmentation – even the simplest segmentation algorithms outperform the subword algorithms largely. Among the subword tokenization algorithms, BPE performed the best on both levels, and the Unigram model performed worst across the board.

There are two major conceptual issues that inhibit a successful transfer of these algorithms to morpheme segmentation. First, these models only operate extremely locally – BPE and WordPiece merge bigrams based on a simple co-occurence metric, and Unigram removes unlikely *n*-grams under the assumption that the distribution of all tokens in the vocabulary is statistically independent. This prevents the models from learning relevant information about longer shared substrings, which is the foundation for all successful morpheme segmentation models. The second, and arguably strongest limiting factor is that it is unclear how to determine when a model should stop. In their intended setting, subword tokenization algorithms are designed to define an expressive vocabulary of a tractable size for downstream NLP applications. Hence, a desired vocabulary size is defined a priori, and the subword vocabulary is continually modified until the predefined size is reached. For BPE and WordPiece, the vocabulary size increases monotonically during that process, while it decreases for Unigram. In this context, vocabulary size refers to the number of unique subwords modeled by the respective tokenizer.

Model	Surface Forms	Underlying Forms
BPE	0.51	0.61
WordPiece	0.38	0.36
Unigram	0.34	0.33

Table 5: Average F_1 scores of subword tokenization algorithms for morphological segmentation.

This training set-up leads to two problems. The first is that the desired vocabulary size must be defined before running the model. For morphological segmentation, the ideal vocabulary size naturally will be the size of the morpheme inventory – but if that is already known, then no automated morphological analysis is required anymore. For the

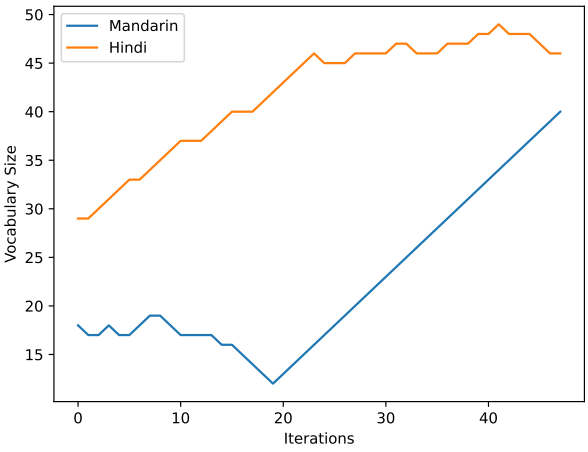


Figure 3: Vocabulary size of the BPE tokenizer for Mandarin and Hindi after each iteration.

sake of illustration, we ran the algorithms under the unrealistic assumption that the ideal vocabulary size is already known; so that each model stopped the training routine once that size was reached. The numbers shown in Table 5 therefore report the performance of an ideal setting for the models, including information that would be unknown in a practical application. BPE and WordPiece reached that ideal vocabulary size only in 11 out of 100 cases (and even then did not provide an ideal morphological segmentation by any means). An accurate reduction of the vocabulary to its minimal representation was therefore rarely achieved.

The second problem results from the assumption that BPE and WordPiece lead to a monotonic increase of vocabulary size. This assumption does not hold true in the special case of numerals: Thanks to the high degree of compositionality, the smallest possible vocabulary size to construct the data is not necessarily the set of individual characters, but can be the set of employed morphemes instead. This is visualized in Figure 3: The Mandarin numerals only require 10 morphemes to construct numerals up to 40, while 19 distinct segments can be found in these forms. By subsequently merging common bigrams, the BPE algorithm is actually able to *reduce* the vocabulary size to these ten morphemes. The monotonicity assumption implied by subword algorithms therefore might be violated, and the vocabulary size might *decrease* for a while, depending on the complexity of a language's morphology and phonology. However, this is not necessarily the case, as in more opaque languages like Hindi, the vocabulary size still increases monotonically with more iterations.

5 Discussion and Conclusion

In this study, we have demonstrated an efficient, transparent, and robust workflow for the annotation and analysis of numeral systems. The workflow features a detailed annotation scheme for shared morphemes across word forms, accounts for potential allomorphy, and can be carried out in a computer-assisted manner, using a web-based annotation tool. As a result, we presented a small sample of annotated numeral systems from 25 typologically diverse languages from Eurasia and South America. We used this sample to evaluate how well unsupervised methods for automated morpheme segmentation work in extremely low-resource scenarios with an extraordinarily strong morphological signal. The results suggest that the major error source of these models is allomorphy. When this factor is accounted for, rather satisfactory morphological analyses can be inferred automatically. For future research on morpheme segmentation in low-resource scenarios, the handling of allomorphy will therefore be crucial.

Several statistical measures of numeral systems introduced here confirm intuitive correlations, such that smaller morpheme inventories necessarily entail a higher expressivity of the individual morphemes. It remains unclear, however, if a measure of morphological complexity can be inferred from our measures, since information-theoretic approaches that have been proposed to measure morphological complexity on corpus data do not convey any useful information about the morphological structure of numeral systems. Curiously, it seems that the performance of Morfessor aligns with (impressionistic) human judgement of how transparent a numeral system is. Since Morfessor is based on the Minimum Description Length (MDL, Rissanen, 1983) principle, which has been proposed as a framework for measuring complexity in numeral systems (Hammarström, 2008; Cathcart, 2025), it might serve as a useful indicator for complexity when applied on the underlying data representation.

We conclude that due to their high degree of compositionality, numerals serve as an ideal controlled sample for developing and testing the annotation and inference of morphological structures in multilingual wordlists. In the future, we hope to further expand our sample of numeral systems and test more methods for automated morpheme segmentation.

Supplementary Material

The dataset for compositional structures in numeral systems (*CoSiNuS*, Version 1.1) is curated on GitHub (https://github.com/numeralbank/cosinus) and archived with Zenodo (https://doi.org/10.5281/zenodo.15656420). The MorSeg software package is curated on GitHub (https://github.com/lingpy/morseg, Version 0.1) and archived with PyPi (https://pypi.org/project/morseg). The code that was used to run the analyses described in this study is curated on Codeberg (https://codeberg.org/calc/numeral-annotation-study, Version 1.1) and archived with Zenodo (https://doi.org/10.5281/zenodo.15672425).

Limitations

The annotation of word forms that etymologically share the same origin, but have diverged over a substantial amount of time, is not always clear and can be ambiguous. For example, consider Spanish *once* (11): There is no transparent, synchronous pattern that would combine *uno* (1) and *diez* (10) to yield this form. However, we know that this was historically the case, as proven by Latin *undecim*, which is a clear compound from *un-* (1) and *decem* (10). In Italian, this compounding strategy is still transparently visible (*un-* + *dieci* = *ùndici*). Arguably, this lexical motivation is still transparent enough in Italian to annotate it as dimorphemic form, but not in Spanish (even though the etymology and the time depth is identical). A similar case can be observed for the Gaelic languages, where the suffix for deriving tens (Irish: *déag*; Scottish: *deug*) is clearly related to the word for ten (*deich* in both languages), but the exact historical connection is unclear (Matasović, 2009, 93-94; MacBain, 1911, 130).

A further limitation to the current annotation scheme is that it linearly segments complex forms into morphemes, for example *two ten three*. The annotation does not make the underlying arithmetic process explicit: Understanding that the underlying formula would be $2 * 10 + 3$, if the word means 'twenty three', requires an additional interpretation step and is not explicitly coded in the annotation scheme.

Due to its relatively small size of 25 languages, the patterns observed in the data might not reflect universal patterns, especially considering the choice of languages. While we tried to include typologically diverse languages, we are aware that our sample is heavily biased towards Indo-European and Sino-Tibetan languages, and that the macroareas of North America, Africa, and Papunesia are not represented at all.

We furthermore observe a heavy bias towards decimal systems, and even those systems that are not primarily decimal contain some decimal structures. It is therefore impossible to systematically analyze different numeral bases beyond some impressionistic analyses. Finally, it remains an open question if (and how) the morphological complexity of a numeral system or a language in general can be measured.

Acknowledgments

This project was supported by the ERC Consolidator Grant ProduSemy (AR, LC, AK, KB, DS, JML; Grant No. 101044282, see https://doi.org/10.3030/101044282), the ERC Synergy Grant QUANTA (CR; Grant No. 951388, see https://doi.org/10.3030/951388), and the Charles University (AS; project GA UK No. 101924, see https://ufal.mff.cuni.cz/node/2690). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

Author Contributions

CR and JML initiated the study and devised the annotation scheme. AR, CR, and JML were responsible for the data management. AR implemented the algorithms for unsupervised morpheme segmentation and subword tokenization (with contributions by AS and JML) and conducted the experiments. AR, LC, KPvD, AK, KB, and DS contributed annotated data. AR, LC, and JML wrote and revised the draft.

References

Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. 2024. Evaluating subword tokenization: Alien subword composition and oov generalization challenge. *arXiv preprint arXiv:2404.13292*.

Andrea Bender and Sieghard Beller. 2012. Nature and culture of finger counting: Diversity and representational effects of an embodied cognitive tool. *Cognition*, 124(2):156–182.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i-Cancho. 2017. The entropy of words – learnability and expressivity across more than 1000 languages. *Entropy*, 19(6).

Frederic Blum, Carlos Barrientos, Johannes Englisch, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, and Johann-Mattis List. 2025. Lexibank 2: precomputed features for large-scale lexical data [version 1; peer review: 3 approved]. *Open Research Europe*, 5(126):1–19.

Hugo Brandt Corstius, editor. 1968. *Grammars for number words*. Reidel, Dordrecht.

Nicolás Brid, Cristina Messineo, and Johann-Mattis List. 2022. A comparative wordlist for the languages of the gran chaco, south america [version 2; peer review: 2 approved]. *Open Research Europe*, 2(90):1–17.

Andreea S. Calude and Annemarie Verkerk. 2016. The typology and diachrony of higher numerals in Indo-European: a phylogenetic comparative study. *Journal of Language Evolution*, 1(2):91–108.

Chundra Cathcart. 2025. Complexity counts: global and local perspectives on Indo-Aryan numeral systems. *arXiv preprint 2505.21510*, pages 1–30.

Eugene Chan. 2024. Numeral systems of the world's languages. `https://lingweb.eva.mpg.de/channumerals/`. Version updated on February 18, 2024.

Çağrı Çöltekin. 2010. Improving Successor Variety for Morphological Segmentation. In *Proceedings of the 20th Meeting of Computational Linguistics in the Netherlands*, volume 16, pages 13–28.

Çağrı Çöltekin and Taraka Rama. 2023. What do complexity measures measure? Correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard*, 9:27–43.

Bernard Comrie. 1997. Some problems in the theory and typology of numeral systems. In Bohumil Palek, editor, *Proceedings of LP'96. Typology: Prototypes, item orderings, and universals. Proceedings of the conference held in Prague August 20–22, 1996*, pages 41–56. Charles University Press, Prague.

Bernard Comrie. 2005. Endangered numeral systems. In Jan Wohlgemut and Tyro Dirksmeyer, editors, *Bedrohte Vielfalt Aspekte des Sprach(en)tods: Aspects of language death*, pages 203–230. Weissensee, Berlin.

Bernard Comrie. 2020. Revisiting Greenberg's "Generalizations about numeral systems" (1978). *Journal of Universal Language*, 21(2):43–84.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology Helsinki.

R. M. W. Dixon. 2012. *Basic linguistic theory, Vol. 3: Further grammatical topics*. Oxford University Press, Oxford.

Robert Forkel. 2024. *CLDFViz. A Python library providing tools to visualize data from CLDF datasets [Software Library, Version 1.3.0]*. Zenodo, Geneva.

Robert Forkel and Johann-Mattis List. 2020. CLDF-Bench. Give your cross-linguistic data a lift. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, page 6997-7004, Luxembourg. European Language Resources Association (ELRA).

Robert Forkel and Johann-Mattis List. 2024. A new Python library for the manipulation and annotation of linguistic sequences. *Computer-Assisted Language Comparison in Practice*, 7(1):17–23.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, 5(1):1–10.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

John A. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

John A. Goldsmith, Jackson L. Lee, and Aris Xanthos. 2017. Computational learning of morphology. *Annual Review of Linguistics*, 3:85–106.

Robin Gower. 2021. *CSV on the Web*. Swirrl, Stirling.

Joseph H. Greenberg. 1978. Generalizations about numeral systems. In Joseph H. Greenberg, editor, *Universals of Human Language, Vol. 3: Word structure*, pages 249–295. Stanford University Press, Stanford.

Margaret A. Hafer and Stephen F. Weiss. 1974. Word segmentation by letter successor varieties. *Information storage and retrieval*, 10(11-12):371–385.

Harald Hammarström. 2008. Complexity in numeral systems with an investigation into pidgins and creoles. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity. Typology, contact, change*, volume 94 of *Studies in Language Companion Series*. John Benjamins Publishing Company.

Harald Hammarström. 2009. *Unsupervised Learning of Morphology and the Languages of the World*. Ph.D. thesis, Chalmers University of Technology and University of Gothenburg.

Harald Hammarström. 2010. Rarities in numeral systems. In Jan Wohlgemuth and Michael Cysouw, editors, *Rethinking universals: How rarities affect linguistic theory*, volume 45 of *Empirical Approaches to Language Typology*, pages 11–60. Mouton de Gruyter, Berlin.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Harald Hammarström, Martin Haspelmath, Robert Forkel, and Sebastian Bank. 2025. *Glottolog [Dataset, Version 5.2]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Zellig S Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190.

Bernd Heine. 1997. *Cognitive foundations of grammar*. Oxford University Press, New York and Oxford.

Lorenzo Hervás y Panduro. 1786. *Aritmetica delle nazioni e divisione del tempo fra l'Orientali*. Gregorio Biasini, Cesena.

Nathan W. Hill and Johann-Mattis List. 2017. Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting*, 3(1):47–76.

James R. Hurford. 1975. *The linguistic theory of numerals*. Cambridge University Press, Cambridge.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Johann-Mattis List. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pages 9–12, Valencia. Association for Computational Linguistics.

Johann-Mattis List. 2021. Using edictor 2.0 to annotate language-internal cognates in a german wordlist. *Computer-Assisted Language Comparison in Practice*, 4(4).

Johann-Mattis List. 2023. Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology*, 14(1156540):1–10.

Johann-Mattis List. 2025. Productive signs: Towards a computer-assisted analysis of evolutionary, typological, and cognitive dimensions of word families. In David Bradley, Katarzyna Dziubalska-Kołaczyk, Camiel Hamans, Ik-Hwan Lee, and Frieda Steurs, editors, *Contemporary Linguistics: Integrating Languages, Communities, and Technologies*, pages 403–412. Brill, Leiden.

Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. *Cross-Linguistic Transcription Systems. [Dataset, Version 2.1.0]*. Max Planck Institute for the Science of Human History, Jena.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(316):1–31.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*, pages 599–605, Berlin. Association of Computational Linguistics.

Johann-Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon J. Greenhill, and Robert Forkel. 2025a. *CLLD Concepticon [Dataset, Version 3.4.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List and Kellen Parker van Dam. 2024. Computer-assisted language comparison with EDICTOR 3 [invited paper]. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 1–11, Bangkok, Thailand. Association for Computational Linguistics.

Johann-Mattis List, Kellen Parker van Dam, and Frederic Blum. 2025b. *EDICTOR 3. An Interactive Tool for Computer-Assisted Language Comparison [Software Tool, Version 3.1]*. MCL Chair at the University of Passau, Passau.

Alexander MacBain. 1911. *An etymological dictionary of the Gaelic language*. Eneas Mackay, Stirling.

Ranko Matasović. 2009. *Etymological dictionary of Proto-Celtic*. Brill, Leiden, Boston.

Edith A. Moravcsik. 2017. Number. In Alexandra Y. Aikhenvald and R. M. W. Dixon, editors, *The Cambridge handbook of linguistic typology*, pages 440–476. Cambridge University Press, Cambridge.

Michele Pulini and Johann-Mattis List. 2024. Finding language-internal cognates in Old Chinese. *Bulletin of Chinese Linguistics*, 17(1):53–72.

Jorma Rissanen. 1983. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431.

Arne Rubehn and Johann-Mattis List. 2025a. *MorSeg: A Python package for morpheme segmentation in multi- and monolingual wordlists [Software Library, Version 0.1]*. Chair for Multilingual Computational Linguistics, University of Passau.

Arne Rubehn and Johann-Mattis List. 2025b. Partial colexifications improve concept embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association of Computational Linguistics.

Christoph Rzymski, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(13):1–12.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, Kyoto, Japan. IEEE.

Nathanael E. Schweikhard and Johann-Mattis List. 2020. Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics*, 17(1):2–26.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16).

Sebastian Spiegler and Christian Monson. 2010. EMMA: A novel evaluation metric for morphological analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1029–1037, Beijing, China. Coling 2010 Organizing Committee.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. In *Aalto University publication series SCIENCE + TECHNOLOGY*, 25. Aalto University, Helsinki, Finland.

Sami Virpioja, Ville T Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

A Statistics for Individual Languages

Language	Morph.	Expressivity	Opacity	Length	Morfessor	LSPE	Affix
Maltese	25 / 12	4.24 / 8.83	2.08	2.65	0.71 / 1.00	0.64 / 0.84	0.71 / 0.84
Mapudungun	10 / 10	8.80 / 8.80	1.00	2.20	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
Wayuu	18 / 15	7.26 / 8.71	1.20	3.19	0.90 / 0.90	0.89 / 0.95	0.70 / 0.70
Aymara	12 / 10	10.58 / 12.70	1.20	3.17	0.94 / 1.00	0.67 / 0.65	0.73 / 0.74
Telugu	27 / 12	3.26 / 7.33	2.25	2.20	0.53 / 0.88	0.35 / 0.72	0.56 / 0.99
Czech	17 / 11	5.18 / 8.00	1.55	2.20	0.73 / 1.00	0.86 / 0.82	0.95 / 1.00
Russian	17 / 12	5.65 / 8.00	1.42	2.40	0.62 / 0.91	0.69 / 0.81	0.83 / 0.97
Irish	23 / 14	4.20 / 6.89	1.64	2.41	0.61 / 0.89	0.58 / 0.63	0.87 / 0.99
Scottish G.	17 / 13	6.94 / 9.08	1.31	2.95	0.64 / 0.66	0.76 / 0.90	0.61 / 0.81
German	20 / 15	5.20 / 6.93	1.33	2.60	0.81 / 0.81	0.65 / 0.95	0.80 / 0.83
Assamese	41 / 20	1.66 / 3.40	2.05	1.70	0.64 / 1.00	0.55 / 0.79	0.60 / 0.88
Hindi	48 / 17	1.40 / 3.94	2.82	1.68	0.42 / 1.00	0.46 / 0.95	0.43 / 1.00
Sanskrit	29 / 13	3.14 / 7.00	2.23	2.53	0.66 / 0.70	0.55 / 0.53	0.45 / 0.75
French	24 / 19	3.08 / 3.89	1.26	1.85	0.75 / 0.79	0.73 / 0.80	0.67 / 1.00
Italian	27 / 14	3.07 / 5.93	1.93	2.08	0.75 / 0.82	0.67 / 0.77	0.79 / 0.96
Latin	23 / 15	4.00 / 6.13	1.53	2.30	0.73 / 0.82	0.71 / 0.79	0.65 / 0.86
Spanish	25 / 19	3.92 / 5.16	1.32	2.45	0.55 / 0.87	0.82 / 0.89	0.73 / 0.97
Cavineña	13 / 10	10.46 / 13.60	1.30	3.40	0.93 / 1.00	0.46 / 0.57	0.67 / 0.68
H. Quechua	10 / 10	8.80 / 8.80	1.00	2.20	1.00 / 1.00	0.89 / 0.89	1.00 / 1.00
Lamjung Y.	26 / 11	3.58 / 8.45	2.59	2.36	0.69 / 0.89	0.89 / 1.0	0.71 / 0.89
Uipo (M.)	18 / 18	8.50 / 8.50	1.00	3.83	0.40 / 0.40	0.93 / 0.93	0.40 / 0.40
Makyam	27 / 18	4.81 / 7.22	1.50	3.25	0.70 / 0.85	0.70 / 0.71	0.45 / 0.77
Mandarin	10 / 10	8.80 / 8.80	1.00	2.20	1.00 / 1.00	1.00 / 1.00	0.96 / 0.96
Shanghainese	12 / 11	6.50 / 7.09	1.09	1.95	1.00 / 1.00	0.87 / 0.87	1.00 / 1.00
Par. Guaraní	11 / 7	9.55 / 15.00	1.57	2.62	0.91 / 1.00	0.89 / 1.00	0.99 / 1.00

Table 6: Overview of statistics about the different numeral systems for each individual language. Whenever two values are given, the left refers to the surface forms, and the right to the underlying form. Morph. indicates the number of distinct morph(eme)s in the given language. The three rightmost columns indicate the performance of automated morpheme segmentation models in terms of F_1.

Beyond the Data: The Impact of Annotation Inconsistencies in UD Treebanks on Typological Universals and Complexity Assessment

Antoni Brosa-Rodríguez and **M. Dolores Jiménez-López**
Universitat Rovira i Virgili
Avda. Catalunya 35, 43002, Tarragona (Spain)
{antoni.brosa, mariadolores.jimenez}@urv.cat

Abstract

This study explores the impact of annotation inconsistencies in Universal Dependencies (UD) treebanks on typological research in computational linguistics. UD provides a standardized framework for cross-linguistic annotation, facilitating large-scale empirical studies on linguistic diversity and universals. However, despite rigorous guidelines, annotation inconsistencies persist across treebanks. The objective of this paper is to assess how these inconsistencies affect typological universals, linguistic descriptions, and complexity metrics. We analyze systematic annotation errors in multiple UD treebanks, focusing on morphological features. Case studies on Spanish and Dutch demonstrate how differing annotation decisions within the same language create contradictory typological profiles. We classify the errors into two main categories: overgeneration errors (features incorrectly annotated, since do not actually exist in a language) and data omission errors (inconsistent or incomplete annotation of features that do exist). Our results show that these inconsistencies significantly distort typological analyses, leading to false generalizations and miscalculations of linguistic complexity. We propose methodological safeguards for typological research using UD data. Our findings highlight the need for methodological improvements to ensure more reliable cross-linguistic generalizations in computational typology.

1 Introduction

Multilingual corpora with consistent annotation schemes have become invaluable resources for typological research in computational linguistics (O'Horan et al., 2016; Ponti et al., 2019). Among these, Universal Dependencies (UD) (Nivre et al., 2023) stands out as one of the most comprehensive collections of consistently annotated treebanks across diverse languages. The standardized annotation framework of UD has enabled researchers to conduct large-scale cross-linguistic comparisons and formulate typological universals based on empirical data rather than theoretical assumptions (Brosa-Rodríguez and Jiménez-López, 2023; Gerdes et al., 2019). This development has significantly advanced our understanding of linguistic diversity and universals.

However, the promise of consistent cross-linguistic annotation faces substantial challenges in practice. Despite rigorous guidelines and quality control measures, inconsistencies and errors in annotation persist across different treebanks, even within the same language. These inconsistencies, while perhaps minor when considering individual treebanks in isolation, can have significant implications when aggregated for typological studies, potentially leading to incorrect characterizations of languages and flawed formulations of linguistic universals. Our research identified several systematic annotation errors across multiple UD treebanks that directly impact typological characterizations based on UD data.

This paper examines how these annotation inconsistencies affect the formulation of typological universals, description of languages or information regarding linguistic complexity (Brosa-Rodríguez et al., 2024) with a particular focus on morphological features such as gender, number, and verbal mode/tense. We establish a correlation between the concepts of linguistic complexity and linguistic universals. We understand the concept of complexity in terms of the difficulty of learning one language from another (second language acquisition); we interpret universals as structures/categories present in all languages. From this standpoint, we establish an inversely proportional relationship between the two concepts: The greater the degree of shared characteristics between two languages, the less challenging it will be to learn one from the other. In essence, the higher the universality of a language, the lower its complexity level when learned as a second language. Given the interrelationship between typo-

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 43–51
August 1, 2025 ©2025 Association for Computational Linguistics

logical universals and complexity, and considering that typological universals are calculated from treebanks, eliminating inconsistencies in UD treebanks is crucial for accurately calculating linguistic complexity, as these inconsistencies can distort typological profiles, affecting the relationship between universals and complexity, particularly when measuring the ease of learning languages based on their shared features.

We exemplify our research with cases from Spanish and Dutch treebanks to demonstrate how annotation decisions in one treebank can differ substantially from another for the same language, creating contradictory typological profiles. Furthermore, we explore how conversion processes from legacy annotation schemes to UD can introduce systematic biases if not carefully supervised.

The research questions guiding this investigation are: (1) How do annotation inconsistencies in UD treebanks affect typological characterizations of languages? (2) What methodological safeguards can researchers implement to account for these biases when conducting typological studies using UD data?

By addressing these questions, we aim to strengthen the foundation of computational typology while acknowledging the inherent challenges in creating truly consistent cross-linguistic annotation schemes. Rather than diminishing the value of resources like UD, our goal is to enhance their utility by promoting awareness of potential biases and suggesting practical approaches to mitigate their effects on typological research.

2 Theoretical Framework

UD (Nivre et al., 2023) has established itself as a standard framework for syntactic annotation across languages (Marneffe et al., 2013; Zeman, 2008; Petrov et al., 2012), with its primary goal being to capture linguistic universals while accommodating language-specific phenomena. The standardized annotation schema enables cross-linguistic comparison and facilitates typological research on an unprecedented scale in corpora (Haspelmath, 2010). However, the application of a universal schema to typologically diverse languages inevitably creates tension between universal applicability and language-specific accuracy.

The challenges of cross-linguistic annotation have been documented in the literature (Kahane et al., 2021; Gerdes et al., 2018, 2022; Yan and Liu, 2022; Osborne and Gerdes, 2019). These challenges include the difficulty of establishing truly universal categories, the problem of forcing language-specific phenomena into universal frameworks, and the lack of correspondence between UD anotation guidelines and classical linguistic claims or theories. While UD has made significant progress in addressing these issues through detailed guidelines and collaborative development, other authors have proposed alternative proposals in order to enhance these detected problems (Gerdes et al., 2022).

In particular, morphological features present unique challenges for cross-linguistic annotation. Features such as gender or number vary significantly across languages, both in terms of their existence and their manifestation. UD addresses this variability through a feature inventory that distinguishes universal from language-specific features. Even if the annotation scheme is adaptable enough, the problems still arise due to annotators (or annotating) action. In this case we do not find as much error analysis as in the case of the revision of the annotation scheme from a theoretical perspective (Arista, 2022; Oh et al., 2020). The only frequent review is a specific review of problems inherent to certain languages, without being general or extendable.

3 Typology of Annotation Errors

Based on our analysis of UD treebanks, we propose a typology of annotation errors that affects typological generalizations. These errors can be classified into two broad categories:

1. **Overgeneration errors**: These occur when features that do not exist in a language (or structure) are incorrectly annotated. We have identified two primary sources of overgeneration:
    - *Automatic conversion artifacts*: When non UD-native treebanks are converted from legacy annotation schemes to UD one, features may be erroneously carried over or generated based on superficial similarities with other languages or parts of speech.
    - *Overgeneralization of specific contexts*: Annotators may apply features appropriate in one context and these are (probably automatically) propagated to con-

texts where they are linguistically unmotivated.

2. **Data omission errors**: These occur when features that do exist in a language (or structure) are inconsistently or incompletely annotated. Sources include:
    - *Annotation fatigue*: Manual annotation of features that are not morphologically marked may be inconsistent due to human error or oversight.
    - *Implicit vs. explicit marking*: Disagreement among annotators regarding whether features should be annotated only when explicitly marked or also when implicitly present through other patterns.

3.1 Implications for Typological Universals

These annotation inconsistencies have direct implications for the identification and validation of typological universals and, even, linguistic (structural) complexity. In the context of UD-based typology, we will use as example a revisitation Greenberg (1963) universals.

Specifically, we examine how annotation errors may affect the validity of linguistic type knowledge based on three universals we select for exemplifying:

- **Universal 30**: If the verb has categories of person-number or if it has categories of gender, it always has tense-mode categories.
- **Universal 31**: If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender.
- **Universal 42**: All languages have pronominal categories involving at least three persons and two numbers.

We consider that both types of errors—overgeneration and data omission—can artificially strengthen or weaken the evidence for these universals. Overgeneration errors may create false examples supporting a universal, while data omission may obscure examples that would contradict it. The combined effect can significantly distort our understanding of cross-linguistic patterns or how we can characterize the studied languages.

In the following sections, we present empirical evidence of these error types from Spanish and Dutch treebanks and demonstrate their impact on the universals listed above.

4 Methodology

Our investigation of annotation inconsistencies in UD treebanks follows a systematic methodology designed to identify, categorize, and assess the impact of annotation errors on typological generalizations. This section describes our data selection, query methods, and analytical approach. We analyzed all available treebanks from UD (version 2.15) quering information contained in Greenberg (1963) universals.

4.1 Query Methodology

To systematically identify annotation inconsistencies, we utilized Grew-Match (Guillaume, 2021), a query tool specifically designed for UD treebanks. Grew-Match allows for precise pattern matching across morphosyntactic features and dependencies, making it ideal for cross-treebank comparison.

We formulated targeted queries to detect potential annotation errors related to the universals under investigation. For example, some of the formalisations used in connection with Greenberg universals that have allowed us to uncover errors are:

```
pattern {A[upos=ADJ,!Gender]}
pattern {A[upos=VERB,Gender=Masc]}
pattern {A[upos=PRON, Person=1, !Number]}
```

The first pattern identifies adjectives lacking gender feature, which may indicate data omission errors relevant to Universal 31. The second pattern identifies verbs with masculine gender in agreement with masculine nominal subjects, which may represent overgeneration errors affecting Universal 31. The third pattern locates first-person pronouns without number annotation, potentially impacting Universal 42.

For each query, we:

1. Executed the pattern across all selected treebanks.
2. Counted matches to quantify the prevalence of each pattern.
3. Extracted contextual examples for qualitative analysis.
4. Compared results across different treebanks of the same language.

4.2 Analytical Framework

Our analysis proceeded in two stages:

Stage 1: Identification of Candidate Errors

We first identified candidate errors by looking for patterns that: (1) appeared inconsistently across different treebanks of the same language; (2) contradicted known typological features of the language; (3) showed signs of automatic conversion artifacts, such as systematic misapplication of features.

Stage 2: Impact Assessment

Then, we assessed the impact of confirmed errors on typological universals by: (1) quantifying how the error affects statistical generalizations; (2) determining whether the error would lead to misclassification of a language with respect to a universal; (3) estimating the potential cascade effect on related typological claims.

4.3 Reproducibility

To ensure reproducibility of our findings, we provide all Grew-Match queries used in our analysis. They are available in GitHub.

5 Case Studies

This section presents detailed analyses of specific annotation inconsistencies identified in our investigation and their implications for typological research. We focus on three representative cases that illustrate both overgeneration and data omission errors across different morphological features.

5.1 Gender in Spanish Verbs: An Overgeneration Error

Our analysis revealed a systematic overgeneration error in Spanish treebanks, where perfect participles in compound verb forms are incorrectly annotated with gender features. For example, in the AnCora treebank, sentences like *"Microsoft ha cometido repetidamente graves violaciones legales"* ('Microsoft has repeatedly committed serious legal violations'), show the participle *cometido* annotated with `Gender=Masc`, as can be seen in figure 1.

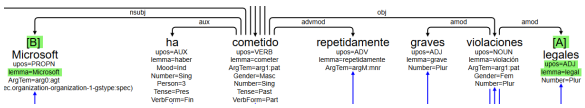


Figure 1: Annotation of "ha cometido" in Spanish-AnCora treebank

This pattern of marking a gender is not limited to AnCora but appears across multiple Spanish treebanks, as evidenced by examples such as *"He dicho con una botella"* ('I have said with a bottle') from COSER, *"Han muerto todos"* ('They have all died') from GSD, and *"Hemos pedido a otros países"* ('We have asked other countries') from PUD.

The error appears to stem from an overgeneralization of specific contexts where gender marking on participles is linguistically motivated (such as in passive constructions like *"fue cometida"* - 'it was committed'). In compound tenses with *haber*, however, the participle functions purely as a verbal element without nominal or adjectival properties, making gender marking inappropriate in these contexts, as Spanish does not express gender in verbs.

Impact on Typological Universals This inconsistency directly affects Universal 31, which concerns patterns of gender agreement. When analyzing Spanish based on these treebanks, we would incorrectly conclude that Spanish exhibits gender marking on verbs in all perfect constructions, potentially classifying it with languages that genuinely mark gender on verbs. Thus, we could also wrongly conclude that there is gender agreement between subjects and verbs. This misclassification could skew cross-linguistic patterns and lead to incorrect typological generalizations about the distribution of gender features across different parts of speech.

5.2 Gender in Roman Languages Adjectives: Implicit vs. Explicit Marking

We identified a systematic data omission problem regarding gender features in invariant adjectives across Spanish treebanks. In the AnCora treebank, noun phrases like *"La admisión oficial"* ('The official admission'), the adjective *oficial* lacks gender annotation, as can be seen in figure 2.

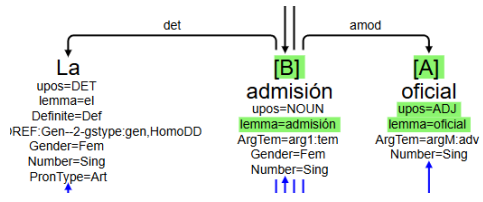


Figure 2: Annotation of "la admisión oficial" in Spanish-AnCora treebank

This contrasts with other Spanish treebanks, which show inconsistent approaches to the same adjective. For example, in GSD we find *"Es el segundo oficial organizado por"* ('It is the second official [event] organized by') with the adjective *oficial* marked as `Gender=Masc`, while in PUD *"Las*

fotos oficiales" (‘The official photos’) shows *oficiales* with `Gender=Fem` annotation, as can be seen in figure 3.

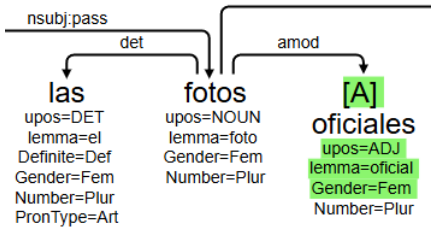


Figure 3: Annotation of "las fotos oficiales" in Spanish-PUD treebank

This inconsistency extends beyond Spanish. Comparing with closely related languages reveals that Portuguese systematically annotates gender on invariant adjectives, as in *"O nome oficial"* (’The official name’) from the Bosque treebank, which includes `Gender=Masc`. Similarly, Italian treebanks show the same inconsistency pattern, with ISDT containing examples of *"ufficiale"* without gender feature while PUD consistently includes the feature.

The omission appears to stem from the lack of overt morphological marking for gender in invariant adjectives like *oficial*, which has the same form for both masculine and feminine. That is why we consider this to be a problem of disparity of annotators, who in an uncoordinated way interpret whether the morphological marking on the adjective takes precedence in order to decide not to mark the gender of that adjective which is in agreement with the noun it modifies.

Impact on Typological Universals This inconsistency affects Universal 31 , which address adjectival agreement patterns. The inconsistent annotations would suggest that Spanish is from one specific type of language depending on the corpus the researcher uses. Additionally, the cross-linguistic inconsistency makes comparative analysis of gender agreement patterns difficult across related Romance languages.

5.3 Number in Dutch Pronouns: A Data Omission Error

Analysis of Dutch treebanks revealed a systematic omission of number features in pronouns across both Alpino and LassySmall treebanks. For example, the first-person plural pronoun *we* (‘we’) in sentences like *"We hebben een concept"* (‘We have a concept’) from Alpino consistently lacks the `Number=Plur` feature. Similarly, the third-person plural pronoun *zij* (‘they’) in *"schamen voor wat zij"* (‘ashamed of what they’) from LassySmall is annotated without number, and the first-person singular *ik* (‘I’) in *"Ik geloof niet"* (‘I don’t believe’) from Alpino lacks the `Number=Sing` feature.

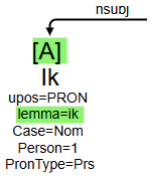


Figure 4: Annotation of "ik" in Duch-Alpino treebank

This pattern extends to all personal pronouns in both Dutch treebanks, creating a systematic gap in the annotation of a fundamental morphological feature. The omission makes it difficult to compare Dutch pronominal systems with those of other languages, where number is consistently annotated on pronouns.

Impact on Typological Universals This fact of not marking number has significant implications for Universal 42, which deals with pronominal number distinctions. Based on Dutch treebanks alone, one might incorrectly conclude that Dutch pronouns lack explicit number marking, placing Dutch typologically with languages that genuinely lack such distinctions. This would represent a substantial mischaracterization of the Dutch pronominal system, which clearly distinguishes singular from plural forms both morphologically and syntactically. Furthermore, the systematic nature of this omission across all pronouns could significantly distort typological comparisons involving pronominal systems.

6 Discussion

Our investigation into annotation inconsistencies in UD treebanks reveals several important implications for typological research. This section examines the broader significance of our findings and proposes approaches to mitigate the impact of annotation errors on typological generalizations.

6.1 Implications for Typological Research

6.1.1 Reliability of Treebank-Based Typology

The systematic errors identified in our case studies raise legitimate concerns about the reliability of typological generalizations derived solely from

treebank data. However, this does not invalidate treebank-based approaches to typology. Rather, it highlights the need for methodological caution when using these resources for cross-linguistic comparison.

Our qualitative analysis suggests that annotation inconsistencies vary across treebanks, with converted resources generally showing more problematic patterns than natively UD-annotated corpora. This is particularly evident in the contrast between AnCora and PUD for Spanish, where PUD exhibits more linguistically motivated annotation of gender on adjectives. This may suggest that typological studies should account for the origin of treebanks when evaluating evidence.

6.1.2 Impact on Specific Universals

Our findings have varying implications for the universals under examination:

Universal 30 (Verbal Features) Universal 30 focuses on verbal tense-mood-aspect systems and is affected by two contrasting error types in Spanish treebanks. First, we observed undergeneration where some verbs receive incomplete tense-mood-aspect annotation while others show complete feature attribution. This inconsistent annotation makes it difficult to accurately characterize the Spanish verbal system in cross-linguistic analysis. Second, we identified an overgeneration problem where verbal participles in perfect constructions are incorrectly assigned gender features. This error conflates verbal and adjectival properties, making Spanish appear to have gender-marking on verbs in contexts where such marking is linguistically unmotivated. Together, these inconsistencies distort the typological classification of Spanish verbal morphology, potentially placing it incorrectly in relation to other languages based on both features it lacks and features it falsely appears to have.

Universal 31 (Gender Agreement) Universal 31 addresses patterns of gender agreement and is significantly impacted by the annotation errors we identified. The spurious assignment of gender to verbs in Spanish compound tenses, observed across multiple treebanks (AnCora, COSER, GSD, PUD), creates the false impression that Spanish typologically aligns with languages that genuinely mark gender on verbs. This overgeneration error artificially expands the scope of gender agreement in Spanish. Conversely, the omission of gender features on invariant adjectives could underrepresent the extent of gender agreement in the language. The cross-linguistic inconsistency in handling invariant adjectives, as seen in our comparison between Spanish, Portuguese, and Italian treebanks, further complicates typological comparisons, as the same linguistic phenomenon receives different treatments across related languages.

Universal 42 (Pronominal Number) Universal 42 concerns pronominal number distinctions and is undermined by the systematic omission of number features on pronouns in Dutch treebanks. Our analysis revealed that both all pronouns, like *we* ('we') or *ik* ('I'), as well as third-person pronouns like *zij* ('they'), consistently lack number annotation in both Alpino and LassySmall treebanks. This pervasive omission could lead to the misclassification of Dutch as having a pronominal system without number distinctions, which would be a fundamental mischaracterization. This is particularly problematic for typological studies that rely on pronoun features to establish diachronic or areal patterns. The systematic nature of this omission across all pronouns in both treebanks suggests a guideline interpretation issue rather than random annotation errors, potentially affecting how Dutch relates typologically to other Germanic and European languages.

6.1.3 Methodological Implications

Our findings suggest that computational typologists should implement several methodological safeguards when working with UD data:

1. **Multi-treebank verification**: When multiple treebanks exist for a language, researchers should compare annotation patterns across resources to identify potential inconsistencies, as demonstrated by our comparison of different Spanish treebanks, even if they are not interested in all textual tipologies.

2. **Conversion awareness**: Studies should explicitly account for whether treebanks were natively annotated in UD or converted from legacy formats, as conversion artifacts represent a significant source of errors.

3. **Cross-linguistic consistency checks**: Researchers should verify whether similar linguistic phenomena receive consistent annotation across related languages, as shown in our comparison of invariant adjectives across Romance languages.

4. **Annotation guideline consultation**: When discrepancies are found, reference to the UD guidelines can help determine which approach better reflects the intended annotation standard.

6.2 Improving UD for Typological Research

While our study identifies several challenges, we believe that UD remains an invaluable resource for computational typology. Based on our findings, we propose several improvements to enhance the reliability of UD for typological research:

6.2.1 Clearer Guidelines for Implicit Features

Many of the data omission errors identified stem from ambiguity regarding whether features should be annotated only when morphologically marked or also when syntactically relevant but not overtly marked. The UD guidelines could be enhanced with more explicit guidance on:

- Annotation of agreement features on invariant forms, as seen in the case of Spanish adjectives.
- Systematic annotation of inherent features on pronouns, as highlighted by the Dutch examples.

6.3 Balancing Universality and Accuracy

The tension between universal application and language-specific accuracy represents a fundamental challenge for cross-linguistic annotation projects. Our case studies illustrate how this tension can manifest in specific annotation decisions, such as whether to annotate gender on invariant adjectives or number on pronouns.

7 Conclusion

This study has identified and analyzed systematic annotation inconsistencies in UD treebanks that affect typological generalizations, with a focus on exemplifying it by morphological features in Spanish and Dutch. Our investigation revealed two primary categories of errors: overgeneration, where features are incorrectly applied to elements that should not have them, and data omission, where features are inconsistently or incompletely annotated. These errors have direct implications for the validity of typological universals derived from UD data.

7.1 Summary of Findings

Our case studies demonstrated specific instances of annotation inconsistencies with typological consequences:

- Incorrect assignment of gender features to verbal participles in Spanish compound tenses across multiple treebanks, creating a false impression that Spanish verbs carry gender marking.
- Inconsistent annotation of gender on invariant adjectives across Spanish treebanks, creating artificial variation within the same language.
- Cross-linguistic inconsistency in handling invariant adjectives across Romance languages.
- Systematic omission of number features on pronouns in Dutch treebanks, potentially leading to incorrect characterization of Dutch pronominal number distinctions.

7.2 Implications for Typology and Linguistic Complexity

These annotation inconsistencies significantly impact both typological research and linguistic complexity studies. As Brosa-Rodríguez et al. (2024) state, the relationship between typological universals and linguistic complexity is inversely proportional—languages sharing more universal features are generally considered less complex to learn as second languages.

The inconsistencies we identified distort complexity metrics by artificially inflating or deflating the morphological complexity of language systems. For instance, spurious gender assignments to Spanish verbs increase the apparent verbal complexity, while omitted number features in Dutch pronouns potentially underestimate pronominal complexity. Such distortions compromise cross-linguistic comparisons and may lead to incorrect predictions about second language acquisition challenges.

These implications underscore the need for researchers to carefully account for annotation inconsistencies when using UD data for both typological research and complexity measurements.

7.3 Contributions

This research makes several contributions to the field of computational typology:

- A typology of annotation errors that affect typological generalizations.

- Empirical evidence of specific inconsistencies in widely used UD treebanks.
- Methodological recommendations for typological research using UD.

7.4 Future Directions

Building on our findings, several promising directions for future research emerge:

- Development of validation procedures to identify typologically relevant annotation inconsistencies.
- Expansion of this analysis to other languages and language families.
- Investigation of how annotation inconsistencies affect typological metrics, language classification, and complexity measurements.
- Collaboration with the UD community to refine annotation guidelines.

7.5 Final Remarks

Despite the challenges identified, we remain optimistic about the value of UD for typological research and complexity studies. By acknowledging and addressing annotation inconsistencies, the computational linguistics community can enhance the reliability of treebank-based analyses, ultimately leading to more accurate characterizations of linguistic diversity, universals, and complexity. As multilingual NLP advances, improved consistency in linguistic annotations will strengthen both our theoretical understanding and the foundation for truly multilingual language technologies.

8 Limitations

While our study provides valuable insights into annotation inconsistencies in UD treebanks, several limitations should be acknowledged.

Our investigation relied primarily on an initial explorative qualitative analysis of specific examples rather than comprehensive quantitative assessment. This approach allowed for detailed linguistic analysis but limits our ability to make broad generalizations about the overall prevalence of these inconsistencies across UD treebanks.

The study focused on exemplifying in Spanish, Italian, Portuguese, and Dutch, Indo-European languages with similar typological profiles. This limited language sample may not capture the full range of annotation challenges present across typologically diverse languages. Additionally, our analysis concentrated on morphology, leaving other syntactic features unexplored.

We have theorized about potential effects on universals 30, 31, and 42, but, due to lack of space, we have not empirically validated how correction of these errors would alter cross-linguistic generalizations in practice. This makes it difficult to assess the practical significance of these inconsistencies for typological research.

Our study offers limited insight into the underlying causes of these inconsistencies beyond the broad distinction between conversion artifacts and manual annotation variability. A more detailed understanding of annotation decision processes would provide valuable context for addressing these issues.

Finally, in some cases, multiple theoretically justified annotation approaches may exist for certain features. We did not systematically explore where annotation differences might reflect legitimate theoretical disagreements rather than errors, nor did we propose mechanisms for accommodating such variation within the UD framework.

References

Javier Martín Arista. 2022. Toward the morpho-syntactic annotation of an old english corpus with universal dependencies. *Revista de Linguistica y Lenguas Aplicadas*, 17:85–97.

Antoni Brosa-Rodríguez, M. Dolores Jiménez-López, and Adrià Torrens-Urrutia. 2024. Exploring the complexity of natural languages: A fuzzy evaluative perspective on greenberg universals. *AIMS Mathematics*, 9:2181–2214.

Antoni Brosa-Rodríguez and María Dolores Jiménez-López. 2023. A typometrical study of greenberg's linguistic universal 1. In *Distributed Computing and Artificial Intelligence. Lecture Notes in Networks and Systems*, pages 186–196. Springer.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. Sud or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to ud. In *Proceedings of the Second Workshop on Universal Dependencies*, pages 66–74. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2022. Starting a new treebank? go sud! theoretical and practical benefits of the surface-syntactic distributional approach. In *SyntaxFest Depling 2021 - 6th International Conference on Dependency Linguistics*, pages 35–46.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2019. Rediscovering greenberg's word order universals in ud. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 124–131.

Joseph H. Greenberg. 1963. *Universals of Language*. The M.I.T. Press. Citado intro.

Bruno Guillaume. 2021. Graph matching and graph rewriting: Grew tools for corpus exploration, maintenance and conversion. In *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics*, pages 1–9.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86:663–687.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of ud and sud treebanks for spoken corpora: a proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories*.

Marie-Catherine De Marneffe, Miriam Connor, Natalia Silveira, Samuel R Bowman, Timothy Dozat, and Christopher D Manning. 2013. More constructions, more genres: Extending stanford dependencies. In *Proceedings of the Second International Conference on Dependency Linguistics*, pages 187–196.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2023. Universal dependencies.

Tae Hwan Oh, Ji Yoon Han, Hyonsu Choe, Seokwon Park, Han He, Jinho D. Choi, Na-Rae Han, Jena D. Hwang, and Hansaem Kim. 2020. Analysis of the penn korean universal dependency treebank (pkt-ud): Manual revision to build robust parsing model in korean. In *Ithaca arXiv*.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308.

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of universal dependencies (ud). *Glossa*, 4.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the conference on Language Resources and Evaluation*, pages 2089–2096.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45:1–56. Citado intro.

Jianwei Yan and Haitao Liu. 2022. Semantic roles or syntactic functions: The effects of annotation scheme on the results of dependency measures. *Studia Linguistica*, 76:406–428.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of LREC*.

Beyond cognacy

Gerhard Jäger
University of Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

Abstract

Computational phylogenetics has become an established tool in historical linguistics, with many language families now analyzed using likelihood-based inference. However, standard approaches rely on expert-annotated cognate sets, which are sparse, labor-intensive to produce, and limited to individual language families. This paper explores alternatives by comparing the established method to two fully automated methods that extract phylogenetic signal directly from lexical data. One uses automatic cognate clustering with unigram/concept features; the other applies multiple sequence alignment (MSA) derived from a pair-hidden Markov model. Both are evaluated against expert classifications from Glottolog and typological data from Grambank. Also, the intrinsic strengths of the phylogenetic signal in the characters are compared. Results show that MSA-based inference yields trees more consistent with linguistic classifications, better predicts typological variation, and provides a clearer phylogenetic signal, suggesting it as a promising, scalable alternative to traditional cognate-based methods. This opens new avenues for global-scale language phylogenies beyond expert annotation bottlenecks.

1 Introduction

Originally developed in computational biology, quantitative methods for phylogenetic reconstruction using likelihood-based inference frameworks have now gained widespread acceptance in comparative linguistics. This is evident from the growing number of computational phylogenies proposed for some of the world's largest language families, including Dravidian (Kolipakam et al., 2018), Sino-Tibetan (Sagart et al., 2019), and Indo-European (Heggarty et al., 2023). Moreover, fully automated approaches — where even cognate identification is performed algorithmically — have demonstrated a surprising degree of robustness (Rama et al., 2018). In contrast to the pre-computational era of historical linguistics, where such detailed reconstructions were rare, the generation of fully resolved phylogenies with branch lengths and, in some cases, estimated divergence dates has now become a standard practice in studies of language evolution.

Despite the increasing recognition of computational language phylogenies as a useful addition to the comparative linguistics toolkit, skepticism remains prevalent. A key concern raised by critics is that phylogenetic analyses are often based on cognate sets—groups of historically related words—extracted from semantically aligned word lists. Since these cognate sets are based on expert annotations, they are sparse, labor-intensive to acquire, and raise concerns regarding replicability.

Another limitation of phylogenetic inference based on cognate classes is that it is by definition constrained to individual language families. There is legitimate interest in automatically inferred trees spanning larger collections of languages, perhaps from the entire world. Such trees provide information about the strength of evidence for putative macro-families (Jäger, 2015; Akavarapu and Bhattacharya, 2024). Furthermore, they are useful for downstream tasks such as the statistical modeling of global language evolution (Bentz et al., 2018; Bouckaert et al., 2022).

The literature contains several proposed workflows for extracting character matrices from word lists without cognate annotations, which can then used as input for likelihood-based phylogenetic inference. This paper presents a comparison of cognate-based phylogenetic inference with two such proposals, the one by Jäger (2018) and the one by Akavarapu and Bhattacharya (2024). These methods are evaluated in three ways: (1) by comparing the inferred phylogenies with the Glottolog expert classification (Hammarström et al., 2024),

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 52–60
August 1, 2025 ©2025 Association for Computational Linguistics

(2) how well the inferred phylogenies fit to the typological features from (Skirgård et al., 2023), and (3) an estimation of the strength of the phylogenetic signal in the data, which is inferred with the software *PyPythia* (Haag et al., 2022).

2 Materials and Methods

2.1 Materials

Word lists were obtained from Lexibank[1]; List et al. 2023). These datasets contain lexical entries, including the language they belong to, their meaning, form in IPA transcription, and often a manual cognate annotation. The datasets are curated by the Lexibank community and are available in a standardized format, which makes them suitable for computational analyses.

In a first step, 135 Lexibank dataset were selected. In total, this amounts to 2,486,845 lexical entries from 6,845 languages (identified by glottocodes).

For the purpose of evaluation, typological features were obtained from Grambank[2]. This results in 355,097 binary entries from 2,467 languages and 195 typological features.

A subset of Lexibank data was selected according to the following criteria:

- The entry comes from a language with a Glottocode that is present in the Grambank data.
- The entry has an entry for its meaning (`Concepticon_Gloss`) and a manual cognate annotation (`Cognateset_ID`).
- The meaning comes from the 110 concepts with the largest coverage.

This leaves 113,671 entries from 928 languages. For further processing, the IPA transcriptions were converted to the ASJP alphabet using the python package *lingpy* (List and Forkel, 2024).

Constraining the Grambank data to these 928 languages leaves 138,878 binary data points from all 195 features.

The gold standard tree was obtained from Glottolog.[3]

[1] https://github.com/lexibank

[2] https://github.com/grambank/grambank

[3] https://zenodo.org/records/10804582/files/glottolog/glottolog-cldf-v5.0.zip

2.2 Methods

The overall workflow consists of the following steps:

1. Phylogenetic inference
   (a) Generate a binary character matrix from the Lexibank data.
   (b) Infer a phylogenetic tree from this character matrix.
2. Evaluation
   (a) Compare the inferred phylogenetic tree with the Glottolog expert classification.
   (b) Compare the inferred phylogenetic tree with the Grambank typological features.
   (c) Assess the strength of the phylogenetic signal in the data.

Three different methods were used to generate a binary character matrix: (1) binarized expert-annotated cognate classes, (2) a combination of automatic cognate clustering and unigram/concept features as described in Jäger (2018), and (3) a variant of the method developed by Akavarapu and Bhattacharya (2024) using multiple sequence alignment.

2.2.1 Expert-annotated cognate classes (cc)

Here we use the method introduced by Ringe et al. (2002) and Gray and Atkinson (2003). Each cognate class is treated as a character. A language is coded as 1 if it has a cognate in the class, 0 if it has a different cognate class for the same concept, and missing if it has no cognate for the concept. This results in a matrix with 928 rows and 25,913 columns.

Since each cognate class is, by definition, confined to a single language family, this character matrix contains no signal beyond the family level.

In the tables and figures below, this method is referred to as cc (for cognate classes).

2.2.2 Automatic cognate clustering and unigram/concept features (PMI)

The workflow proposed by Jäger (2018) was replicated. This approach uses two types of characters.

- Binarized cognate classes obtained via automatic cognate clustering. This involves (1) supervised training of a Support Vector Machine classifier which takes a pair of words

and predicts the labels 1 (cognate) or 0 (non-cognate), using manual cognate classification for supervision, (2) creating a distance matrix for all entries for a given concept from the 100 concepts defined above, and (3) clustering the distance matrix using the *label propagation* algorithm (Raghavan et al., 2007).

- Unigram/concept characters. For each combination of a concept c and an ASJP sound class s, a language is coded as 1 if it has a word for concept c that contains sound class s, missing if it has no word for concept c, and 0 otherwise.

This resulted in a matrix with 928 rows and 41,013 columns.

Since the *pointwise mutual information* between sound classes plays an essential role in this workflow, the method is referred to as PMI.

2.2.3 Multiple sequence alignment (MSA)

The method by Akavarapu and Bhattacharya (2024) was used as starting point, but the present approach differs in various aspects. The method is based on the following steps:

In a first step, pairwise distances between languages in the full lexibank dataset were computed using the Levenshtein distance on the ASJP transcriptions and aggregating according to the method described in (Jäger, 2018). Language pairs with a distance below 0.7 were considered as *probably related*, using the same heuristics as Jäger (2018). There are 172,681 such language pairs. All word pairs from such a language pair sharing their meaning are treated as *potential cognates*. There are 90,565,486 such word pairs. An equal number of random word pairs were sampled as probable non-cognates. Potential cognates were assigned the label 1 and probable non-cognate the label 0.

In a second step, a classifier was trained on the potential cognates and non-cognates. The classifier consists of a *pair-Hidden Markov Model* (pHMM) (Durbin et al., 1998) and a logistic-regression layer. The classifier was trained for one epoch using the Adam optimizer. The resulting parameters of the pHMM were used in the next step.

A pHMM defines a probability distribution over pairs of aligned sequences of sound classes. This involves (1) emission probabilities for all pairs of sound classes that are matched in the alignment, (2) emission probabilities for individual sound classes if they are aligned with a gap, and (3) transition probabilities between the hidden states *match*, *gap in string 1*, *gap in string 2*, and *final state*.

It is instructive to inspect the emission probabilities in the trained model. In Table 1 the ten sound classes with the highest probability of being matched with /p/ are shown for illustration, together with their log-probabilities. This ranking is in good agreement with linguistic intuitions about potential sound correspondences.

Sound class	Log-probability
p	−2.39
f	−16.35
b	−18.85
v	−23.26
h	−24.03
L	−25.11
g	−27.67
7	−29.74
C	−29.95
I	−30.78

Table 1: The ten sound classes with the highest probability of being matched with /p/ in the trained pHMM, along with their log-probabilities.

Sound class	Log-probability
c	-0.86
j	-1.17
L	-1.54
1	-2.94
I	-8.06
h	-9.37
7	-9.60
i	-10.14
y	-10.24
T	-10.33

Table 2: The ten sound classes with the highest probability of being matched with a gap in the trained pHMM, along with their log-probabilities.

A high probability here is to be interpreted as a high likelihood that instances of these sound classes participate either in insertion or deletion.

The trained pHMM assigns a probability to each pair of aligned sequences. Via the forward algorithm, the probability of a pair of sequences is computed as the sum of the probabilities of all possible alignments between these sequences.

Following Durbin et al. (1998), a null-model

was trained additionally that assigns individual probabilities to both sequences, disregarding the order of sound classes. The log-odds ratio of a pair of words of being generated by the pHMM vs. the null model can interpreted as a measure of the similarity of the two words.

To illustrate this, a collection of ten words were chosen at random from the dataset which all have an edit distance of 1 to the word *baba*, and their log-odds ratios with respect to *baba* were computed. The results are shown in Table 3.

word	log-odds
babae	98.26
babau	96.31
bIba	95.73
bawa	87.55
zaba	85.51
raba	74.58
maba	73.52
eaba	73.50
xaba	71.78
naba	70.94

Table 3: Ten randomly chosen words with an edit distance of 1 from *baba*, alongside with the predicted log-odds to *baba*.

This ranking illustrates that the log-odds predicted by the trained pHMM are consistent with linguistic intuitions about potential cognacy.

In a third step, the trained pHMM was used in combination with the Viterbi algorithm to obtain pairwise sequence alignments for all synonymous word pairs from different languages within the smaller dataset of 928 languages and 110 concepts.

In a fourth step, the pairwise sequence alignments were aggregated to a *multiple sequence alignment* (MSA) using the *T-Coffee* algorithm (Notredame et al., 2000).

Note that all reflexes of a given concept are aligned within a single MSA, regardless of cognacy. Such an MSA implicitly contains information both about cognacy and about sound correspondences.

An example (for a much smaller dataset) is shown in Table 4 for illustration. These are the reflexes of the concept *louse* from the Tungusic languages in the dataset.[4]

As can be seen from this example, the MSA contains information about cognacy, but also about sound correspondences. For example, a `t` in the first column is a proxy for the cognate class `16_lousen-38`. The sound classes `k` and `q`, on the other hand, both correspond to the cognate class `16_lousen-37`, and they additionally reflect a sound change. In column 4, however, the cognate class `16_lousen-38` is split into two sound classes, `k` and `q`, reflecting a sound change. The presence of a sound class, as opposed to a gap, is a proxy of that cognate class. Put differently, binary characters corresponding to a gap are flipped by switching 0s and 1s.

In a fifth step, the MSA was converted to a binary matrix. Two binarization methods were used simultaneously. For a given column in an MSA, a character was created for the presence of a sound class. For column 4 in Table 4, e.g., this character has value 1 for Nanai, Orok and Ulch, and 0 for the other languages. Additionally, for each sound class type in a column, a different character was created. In the example, there are two such characters, one for `k` and one for `q`. The first has value 1 for Nanai and Orok and 0 otherwise, while the second has value 1 for Ulch and 0 otherwise. Languages for which the data do not contain a reflex for a given concept are coded as missing for all relevant characters. If a language has multiple reflexes for a given concept, the maximum value is chosen.

Applying this workflow to all concepts and concatenating the resulting matrices yields the final character matrix 928 rows and 46,409 columns.

As mentioned above, this workflow builds on the method by Akavarapu and Bhattacharya (2024), but differs in various aspects. The mentioned work (1) uses Dolgopolsky sound classes instead of ASJP, (2) finds the MSA using CLUSTALW2 (Larkin et al., 2007) instead of T-Coffee, and (3) omits the binarization steps, working with a multistate model of evolution for phylogenetic inference.

In the tables and figures this method is referred to as MSA.

2.2.4 Phylogenetic inference

We performed phylogenetic inference using *raxml-ng* (Kozlov et al., 2019), which implements maximum-likelihood estimation. The GTR+G model (generalized time-reversible model with

[4]The data are taken from `https://zenodo.org/records/13163376`, which is based on (Oskolskaya et al., 2021).

Language	Cognateset_ID	1	2	3	4	5	6	7	8
Even	16_lousen-37	k	-	u	-	m	-	k	e
Kilen	16_lousen-37	q	h	u	-	m	I	k	I
Negidal	16_lousen-37	k	-	u	-	m	-	k	I
Oroch	16_lousen-37	k	-	u	-	m	-	-	I
Udihe	16_lousen-37	k	-	u	-	m	u	x	I
Nanai	16_lousen-38	t	-	i	k	t	-	-	I
Orok	16_lousen-38	t	-	i	k	t	-	-	I
Ulch	16_lousen-38	t	-	i	q	t	-	-	I

Language	sound class	k	q
Even	0	0	0
Kilen	0	0	0
Negidal	0	0	0
Oroch	0	0	0
Udihe	0	0	0
Nanai	1	1	0
Orok	1	1	0
Ulch	1	0	1

Table 4: Example of a multiple sequence alignment. Alignment cells are shaded to indicate different cognate sets. (left) Binarized version of column 4. (right)

gamma-distributed rates) was used for all analyses. This means that gain rates and loss rates can be different, and that the mutation rates of the different characters can differ but are drawn from the same gamma distribution. The parameters of this distribution are estimated from the data.

Using the standard settings of *raxml-ng*, 20 maximum likelihood tree searches were performed, ten of them starting from random trees and ten from maximum-parsimony trees. The tree with the highest likelihood was chosen as the final result.

2.2.5 Evaluation

Evaluation was conducted on three types of datasets:

- the full dataset of 928 languages,
- 100 samples of 100 languages each, which are drawn at random without replacement from the full dataset, and
- a collection of 14 language families, each with at least 10 languages.

For each of these groups of datasets, the following evaluations were performed:

Comparison with Glottolog The Glottolog classification of the languages in a dataset can be represented as a phylogenetic tree with polytomies, i.e., with nodes containing more than two daughters. This Glottolog tree serves as gold standard. To assess the degree of agreement between the gold standard and the inferred phylogenies, the *generalized quartet distance* (GQD) was deployed, as first proposed by Pompei et al. (2011). This distance is defined as the fraction of quartets (i.e., sets of four languages) that are (a) resolved in both trees, and (b) resolved differently in the two trees. The GQD ranges from 0 (perfect agreement) to 0.67 (chance level). The GQD was computed using the software *QDist*, which can be obtained from https://birc.au.dk/software/qdist/.

Fit with Grambank The hypothesis is assumed that the values of the Grambank features evolve along a phylogeny in the same way as the lexical characters described earlier in this section. The degree of fit of the inferred phylogenies with the Grambank features was assessed by (1) using the inferred phylogeny and estimating the branch lengths, mutation rates and rate heterogeneity via Maximum Likelihood, and (2) computing the *Akaike Information Criterion* (AIC). A lower AIC value indicates a better fit.

ML inference and AIC computation were also performed with *raxml-ng*.

For the groups of random samples and of language families, AIC values were normalized to mean 0 to facilitate comparison.

Phylogenetic difficulty The strength of the phylogenetic signal in the data was assessed using the software *PyPythia* (Haag et al., 2022). The authors define a measure of signal strength that uses 100 maximum likelihood tree searches and quantifies the degree of agreement between the inferred trees. The software *PyPythia* implements a machine learning algorithm that predicts this difficulty from various properties of the character matrix, such as entropy and sites-over-taxa ratio, and maximum-parsimony tree inference, with high precision and comparatively low computational cost. The measure ranges from 0 (little difficulty, i.e., strongest signal) to 1 (very difficult, i.e., no signal).

Method	GQD (Glottolog)	AIC (Grambank)	difficulty
Cognate classes	0.188	105.340	0.59
PMI	0.062	104.903	0.63
MSA	**0.042**	**104,752**	**0.45**

Table 5: Evaluation of the full dataset. GQD = Generalized Quartet Distance (lower is better; ranges from 0 for perfect fit to 0.67 for chance level); AIC = Akaike Information Criterion for typological model fit (lower is better; absolute values are not interpretable in isolation but differences are meaningful); difficulty = phylogenetic difficulty estimated by PyPythia (lower is better; ranges from 0 for strong phylogenetic signal to 1 for absent signal).

Method	μ GQD	σ GQD	μ AIC	σ AIC	μ difficulty	σ difficulty
Cognate classes	0.227	0.077	151	115	0.486	0.030
PMI	0.095	0.030	−28	69	0.326	0.032
MSA	**0.048**	0.015	**−123**	66	**0.294**	0.021

Table 6: Evaluation of the 100 random samples (μ: sample mean; σ: sample standard deviation).

Method	μ GQD	σ GQD	μ AIC	σ AIC	μ difficulty	σ difficulty
Cognate classes	0.223	0.130	**−1.73**	17.01	0.401	0.164
PMI	0.221	0.109	3.42	20.43	0.280	0.187
MSA	**0.218**	0.109	−1.69	14.30	**0.203**	0.159

Table 7: Evaluation of the 14 largest language families (μ: sample mean; σ: sample standard deviation).

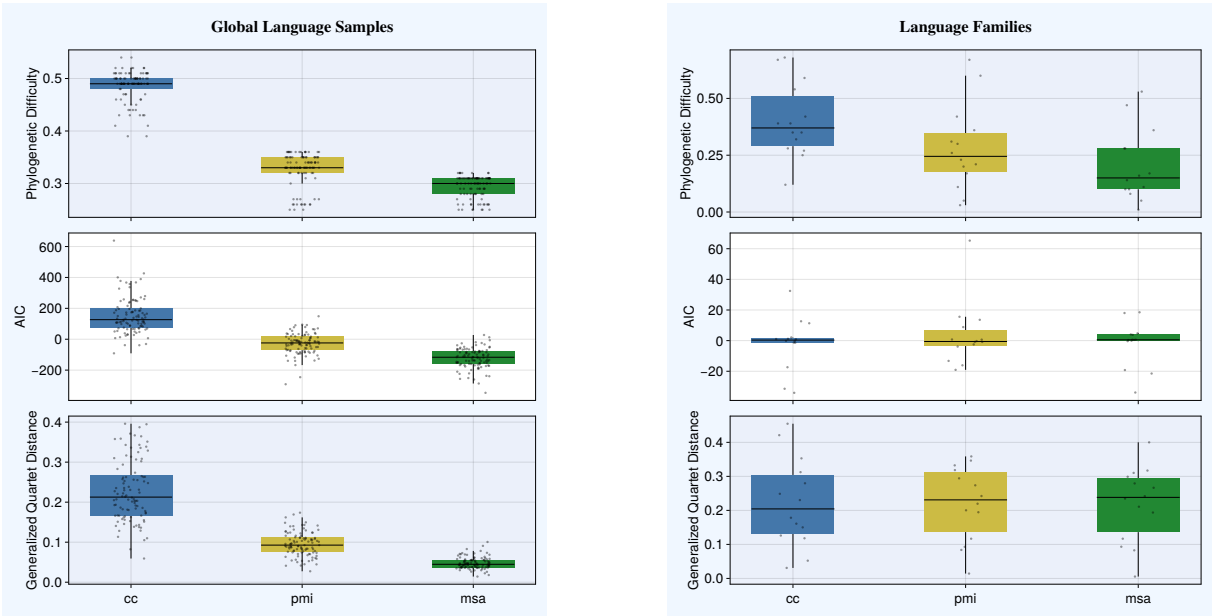


Figure 1: **Left panel:** Comparison of methods across three evaluation metrics for the 100 random samples. The boxplots show distribution per method, while the overlaid points represent individual samples. **Right panel:** Comparison of methods across three evaluation metrics for the 14 largest language families. The boxplots show distribution per method, while the overlaid points represent individual samples.

3 Results

The evaluation results for the entire dataset are shown in Table 5. Table 6 shows the aggregated results for the 100 random samples. They are visualized in the left panel of Figure 1.

The aggregated evaluation results for the 14 largest language families are shown in Table 7. They are visualized in the right panel of Figure 1.

The results for the individual families are given in Table 8.

When focusing on phylogenetic inference at the level of individual families, we find a considerable variation between families. This applies both to the numerical evaluation results and the relative ranking of the three methods considered here. The MSA method tends to produce the lowest phylogenetic difficulty, while there is no discernible trend regarding the fit to Glottolog and to Grambank.

This picture changes considerably when we focus on datasets covering languages from many different families. Here, the MSA method consistently outperforms the other two methods. This is particularly evident in the comparison with Glottolog, where the MSA method yields the lowest GQD values. The MSA method also leads to the lowest AIC values, indicating a better fit to the Grambank typological features. The phylogenetic difficulty is also lowest for the MSA method.

4 Discussion

These findings suggest that the MSA method is a promising alternative to traditional cognate-based methods. It is competitive with the more labor-intensive method based on manual cognate annotations, as well as the method using automatically detected cognate classifications, when considering individual language families. For global datasets, the MSA method clearly outperforms the other two methods. This is particularly evident in the comparison with Glottolog, where the MSA method yields the lowest GQD values. The MSA method also tends to produce the lowest AIC values, indicating a better fit to the Grambank typological features. The phylogenetic difficulty is also lowest for the MSA method.

Limitations

The two evaluation methods that quantify the fit of the inferred trees to empirical data only assess the quality of the inferred tree **topologies**. Future work will need to address the question how well the inferred branch lengths and divergence dates correspond to the true values. This is a challenging task, as the true values are unknown. It is expected that the usefulness for downstream tasks is a suitable proxy.

Data and Code Availability

The code used in this study is available at `https://codeberg.org/profgerhard/sigtyp2025_code/`.

Acknowledgments

This research was supported by the DFG Centre for Advanced Studies in the Humanities Words, Bones, Genes, Tools (DFG-KFG 2237) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 834050).

References

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024. A likelihood ratio test of genetic relationship among languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2559–2570, Mexico City, Mexico. Association for Computational Linguistics.

Christian Bentz, Dan Dediu, Annemarie Verkerk, and Gerhard Jäger. 2018. The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour*, 2(11):816–821.

Remco Bouckaert, David Redding, Oliver Sheehan, Thanos Kyritsis, Russel Gray, Kate E Jones, and Quentin Atkinson. 2022. Global language diversification is linked to socio-ecology and threat status. *SocArXiv*.

Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Russell D Gray and Quentin D Atkinson. 2003. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965):435–439.

Julia Haag, Dimitri Höhler, Ben Bettisworth, and Alexandros Stamatakis. 2022. From easy to hopeless—predicting the difficulty of phylogenetic analyses. *Molecular biology and evolution*, 39(12):msac254.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1. *Leipzig: Max Planck Institute for Evolutionary Anthropology.(Available online atglottolog. org, Accessed on2024-11-29.)*, 10.

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F.

Family	Metric	cc	pmi	msa
Afro-Asiatic	GQD	0.455	**0.195**	0.211
	PhyDiff	0.420	0.260	**0.100**
	AIC	12.635	8.841	**-21.476**
Arawakan	GQD	**0.280**	0.346	0.317
	PhyDiff	0.250	0.110	**0.080**
	AIC	0.868	**-0.845**	-0.023
Atlantic-Congo	GQD	**0.150**	0.220	0.235
	PhyDiff	0.680	0.670	**0.530**
	AIC	**-34.076**	15.572	18.504
Austroasiatic	GQD	**0.052**	0.093	0.093
	PhyDiff	**0.280**	0.310	0.280
	AIC	11.338	**-16.115**	4.777
Austronesian	GQD	**0.161**	0.200	0.194
	PhyDiff	0.670	0.600	**0.470**
	AIC	-31.419	65.299	**-33.880**
Chibchan	GQD	**0.118**	0.332	0.310
	PhyDiff	0.350	0.360	**0.110**
	AIC	-0.168	**-3.895**	4.063
Dravidian	GQD	0.312	0.242	**0.242**
	PhyDiff	0.390	0.170	**0.100**
	AIC	1.196	**-0.862**	-0.334
Indo-European	GQD	0.031	0.014	**0.005**
	PhyDiff	0.320	0.210	**0.140**
	AIC	1.065	**-19.079**	18.015
Pama-Nyungan	GQD	**0.178**	0.359	0.299
	PhyDiff	0.540	0.420	**0.360**
	AIC	**-17.375**	13.601	3.774
Sino-Tibetan	GQD	**0.230**	0.318	0.279
	PhyDiff	0.590	0.300	**0.280**
	AIC	32.455	-13.239	**-19.216**
Tucanoan	GQD	0.421	**0.274**	0.400
	PhyDiff	0.270	0.030	**0.010**
	AIC	**-1.364**	0.758	0.607
Tupian	GQD	0.353	0.294	**0.266**
	PhyDiff	0.390	0.200	**0.160**
	AIC	-0.187	**-0.280**	0.467
Turkic	GQD	0.249	**0.117**	**0.117**
	PhyDiff	0.350	0.230	**0.170**
	AIC	**-1.266**	0.647	0.618
Uto-Aztecan	GQD	0.126	0.084	**0.083**
	PhyDiff	0.120	**0.050**	0.050
	AIC	2.098	**-2.485**	0.388

Table 8: Evaluation of the 14 largest language families. The best value for each family is highlighted in bold.

Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, and 14 others. 2023. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656).

Gerhard Jäger. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, 112(41):12752–12757.

Gerhard Jäger. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5(1):1–16.

Vishnupriya Kolipakam, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco Bouckaert, Russell D. Gray, and Annemarie Verkerk. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, 5(171504):1–17.

Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. 2019. Raxml-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Mark A Larkin, Gordon Blackshields, Nigel P Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, and 1 others. 2007. Clustal w and clustal x version 2.0. *bioinformatics*, 23(21):2947–2948.

Johann-Mattis List and Robert Forkel. 2024. Lingpy. a python library for historical linguistics. With contributions by Simon Greenhill, Tiago Tresoldi, Christoph Rzymski, Gereon Kaiping, Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, Patrick Elmer, Arne Rubehn.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2023. Lexibank analysed. *Scientific Data*, 9(316):1–31. Data set.

Cédric Notredame, Desmond G Higgins, and Jaap Heringa. 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217.

S. Oskolskaya, E. Koile, and M. Robbeets. 2021. A Bayesian approach to the classification of Tungusic languages. *Diachronica*, 39(1):128–158.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PLoS One*, 6(6):e20109.

Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 76(3):036106.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400.

Donald Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Science of the United States of America*, 116:10317–10322.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, and 86 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9.

SenWiCh: Sense-Annotation of Low-Resource Languages for WiC using Hybrid Methods

Roksana Goworek[1,2], Harpal Karlcut[1], Hamza Shezad[1], Nijaguna Darshana[1], Abhishek Mane[1], Syam Bondada[1], Raghav Sikka[1], Ulvi Mammadov [1], Rauf Allahverdiyev[1], Sriram Purighella[1], Paridhi Gupta[1], Muhinyia Ndegwa[1], Bao Khanh Tran[1], Haim Dubossarsky[1,2,3]

[1]Queen Mary University of London, [2]The Alan Turing Institute, [3]University of Cambridge,

Abstract

This paper addresses the critical need for high-quality evaluation datasets in low-resource languages to advance cross-lingual transfer. While cross-lingual transfer offers a key strategy for leveraging multilingual pretraining to expand language technologies to understudied and typologically diverse languages, its effectiveness is dependent on quality and suitable benchmarks. We release new sense-annotated datasets of sentences containing polysemous words, spanning ten low-resource languages across diverse language families and scripts. To facilitate dataset creation, the paper presents a demonstrably beneficial semi-automatic annotation method. The utility of the datasets is demonstrated through Word-in-Context (WiC) formatted experiments that evaluate transfer on these low-resource languages. Results highlight the importance of targeted dataset creation and evaluation for effective polysemy disambiguation in low-resource settings and transfer studies. The released datasets and code aim to support further research into fair, robust, and truly multilingual NLP.

1 Introduction

Cross-lingual transfer is a key strategy in modern NLP, particularly for low-resource languages, where training data is scarce. By leveraging multilingual pretraining, models can transfer task-specific abilities from high-resource languages to low-resource ones, expanding access to language technologies for underrepresented communities (He et al., 2021; Ponti et al., 2018; Wei et al., 2021).

Despite its promise, transfer learning is not universally effective across tasks or languages. Studies on tasks like POS tagging, NER, NLI, QA, and sentiment analysis (Pires et al., 2019; Dolicki and Spanakis, 2021; Srinivasan et al., 2021; Lauscher et al., 2020; Ahuja et al., 2023), as well as polysemy disambiguation (Raganato et al., 2020; Dairkee and Dubossarsky, 2024), show that cross-lingual transfer can be inconsistent and, in some cases, fail entirely. This is also true for generative models (Robinson et al., 2023; Shaham et al., 2024; Chirkova and Nikoulina, 2024), with particularly poor performance in low-resource languages, highlighting the need for more robust and language-inclusive transfer.

A main obstacle for transfer is the lack of high-quality datasets in low-resource and typologically diverse languages. Without these benchmarks, assessing transfer performance, let alone training models on target languages, remains a formidable challenge. This lacking is largely due to the scarcity of linguistic resources in low-resource languages. For instance, Wiktionary contains over a million entries for German, English, French, Chinese, and Russian, but fewer than 100,000 for Punjabi and Marathi (Wikimedia Foundation, 2025).

This lack of resources underscores the urgent need for dedicated datasets to evaluate and refine transfer techniques for underrepresented languages, which this work addresses by developing a semi-automatic method for sense annotation in polysemy and generating resources in ten languages.

We focus on the task of polysemy disambiguation, as it particularly challenges cross-lingual transfer by revealing structural and semantic differences between languages. While some NLP tasks, like sentiment analysis, rely on meaning preservation across languages, where direct translation can maintain performance, polysemy is highly language-specific (Rzymski et al., 2020), making it a rigorous test of a model's ability to generalize across languages. For example, the English word "movement" refers to both physical motion and a political or social movement. However, its Polish translation, "ruch", also encompasses these two meanings, but additionally means "traffic", a sense not covered by the English word. Conversely, "movement" in English can also refer to a section of a musical composition.

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 61–74
August 1, 2025 ©2025 Association for Computational Linguistics

Polysemy disambiguation has long been considered a hallmark of human cognition and a central challenge in NLP (Navigli, 2009; Bevilacqua et al., 2021). A model that can accurately distinguish between different senses of a word must capture linguistic subtleties, metaphorical meanings, and even emerging word usages, much like human speakers. Thus, success in cross-lingual polysemy disambiguation would suggest a model's ability to generalize deep semantic understanding, beyond surface-level patterns in a single language. While many high-resource languages already benefit from sense-annotated datasets (see §2), low-resource languages remain largely unrepresented in this area. Existing contextualized models can process polysemous words within downstream tasks (Loureiro et al., 2021; Ushio et al., 2021), but sense disambiguation remains a major challenge across dozens of languages (Pilehvar and Camacho-Collados, 2019a; Raganato et al., 2020; Martelli et al., 2021; Liu et al., 2021).

Beyond NLP, polysemy also presents difficulties in multimodal models, such as object detection systems, where the same word can refer to multiple visual categories (Calabrese et al., 2020). This suggests that solving polysemy is not just beneficial for language tasks but has broader implications for AI reasoning and multimodal understanding.

Our Contributions Despite extensive work on polysemy disambiguation in high-resource languages, datasets for low-resource languages remain scarce. We address this gap with the following contributions:

- **Sense-annotated datasets:** We release both WSD-style sense-annotated corpora and WiC-style evaluation datasets for ten low-resource languages.[1] The WiC format supports direct comparison with existing experiments in other languages, enabling strong cross-lingual baselines.

- **Annotation tool:** To facilitate further resource development, we release a hybrid semi-automated annotation tool.[2]

Together, these contributions represent a crucial step toward advancing fair, robust, and truly multilingual NLP by enabling evaluation and development in languages that have been largely neglected.

[1]available at DOI: 10.5281/zenodo.15493005

[2]available at github.com/roksanagow/projecting_sentences

2 Related Work

2.1 Transfer Studies

Zero-shot cross-lingual transfer has been widely studied, with mixed findings on its effectiveness, particularly in polysemy disambiguation. While some studies highlight transfer potential across languages, others expose significant limitations, especially in tasks that depend on fine-grained semantic distinctions.

Lauscher et al. (2020) examined zero-shot transfer performance across 17 languages and five NLP tasks (excluding polysemy), evaluating XLM-R (Conneau et al., 2020) and mBERT (AI, 2018). They found that zero-shot performance drops significantly compared to full-shot settings and that transfer success correlates with factors like pretraining corpus size and linguistic similarity. These findings suggest that cross-lingual transfer is far from universal and is highly dependent on language resources and pretraining coverage.

Focusing specifically on polysemy disambiguation, Raganato et al. (2020) conducted the first large-scale cross-lingual transfer study for this task, training a model on English and evaluating on 12 other languages. While they observed some zero-shot transferability, models trained on English underperformed models trained on the target language by 10-20% when tested on German, French, and Italian, indicating that polysemy disambiguation remains language-sensitive and benefits from in-language supervision.

In contrast, Dairkee and Dubossarsky (2024) challenged the feasibility of cross-lingual transfer for polysemy disambiguation altogether. Studying English and Hindi, they found a complete lack of zero-shot transfer, suggesting that word sense distinctions may be too language-specific for direct transfer without explicit in-language supervision.

These conflicting results emphasize the need for more comprehensive transfer studies in polysemy disambiguation, particularly in low-resource languages where transfer learning is often the only viable approach due to the lack of labeled data. However, without high-quality evaluation datasets in these languages, assessing and improving transfer learning for polysemy remains an open challenge.

2.2 Polysemy Disambiguation

Word Sense Disambiguation (WSD) datasets are sense-annotated corpora consisting of sentences containing polysemous words, labeled according

to their contextual meanings. WSD is inherently complex, as words vary in the number of possible senses, and the list of words differs across languages. To address this, Pilehvar and Camacho-Collados (2019a) introduced the Word in Context (WiC) formulation, which reformulated the original WSD problem, which was a multi-class classification task, into a binary classification one. Instead of assigning specific sense labels, WiC pairs two sentences containing the same word and labels them 1 (same) or 0 (different). For example:

> A **bat** flew out of the cave as the sun set.
> He swung the **bat** with all his strength.

This approach enables models to be trained directly on polysemy disambiguation by adjusting embeddings so that words with the same sense cluster together, while those with different senses are pushed apart in the resulting embedding space.

2.3 Existing Datasets

2.3.1 WSD Datasets

Word Sense Disambiguation (WSD) research has been supported by several key sense-annotated corpora and lexical resources:

SemCor (Miller et al., 1993) is a foundational English corpus containing over 226,000 sense annotations across 352 documents.

OntoNotes (Hovy et al., 2006) offers a multi-genre corpus with extensive annotations, including word senses linked to a refined sense inventory for English, Chinese and Arabic.

Senseval/SemEval Datasets have been instrumental in standardizing WSD evaluation. Notably, **Senseval-2** (Edmonds and Cotton, 2001) and **SemEval-2007 Task 17** (Pradhan et al., 2007) provided all-words WSD tasks, challenging systems to disambiguate every content word in given texts. These competitions have included data in multiple languages, such as English, Chinese, Basque, and others (Navigli et al., 2013).

CoarseWSD-20 (Loureiro et al., 2021) is a coarse-grained sense disambiguation dataset derived from Wikipedia, focusing on 20 ambiguous nouns, each with 2 to 5 senses, all in English.

FEWS (Few-shot Examples of Word Senses) (Blevins et al., 2021) addresses the challenge of disambiguating rare senses. Automatically extracted from Wiktionary, FEWS provides a large training set covering numerous senses and an evaluation set with few- and zero-shot examples, facilitating research in low-shot WSD scenarios in English.

WordNet (Miller, 1995) serves as a comprehensive lexical database grouping words into synsets representing distinct concepts. Each synset is interconnected through various semantic relations, offering a structured sense inventory integral to WSD tasks. It primarily focuses on English, but various projects have extended it to other languages.

BabelNet (Navigli and Ponzetto, 2012) extends WordNet by integrating it with Wikipedia and other resources, forming a multilingual semantic network. As of version 5.3 (December 2023), BabelNet covers 600 languages, containing almost 23 million synsets and around 1.7 billion word senses (Navigli et al., 2023). This expansive resource connects concepts across languages, supporting cross-lingual WSD and enriching the sense inventory beyond monolingual constraints.

2.3.2 WiC Datasets

The Word-in-Context (WiC) framework has been instrumental in evaluating context-sensitive word embeddings through binary classification tasks. Several notable datasets have been developed within this framework:

WiC (Pilehvar and Camacho-Collados, 2019b) is the pioneering English dataset that introduced the WiC framework. It consists of sentence pairs where a target word appears in both contexts, and the task is to determine whether the word carries the same meaning in both sentences. This dataset has set the standard for subsequent WiC-based evaluations.

XL-WiC (Raganato et al., 2020) extends the WiC framework to a multilingual setting, encompassing 12 languages: Bulgarian, Danish, German, Estonian, Farsi, French, Croatian, Italian, Japanese, Korean, Dutch, and Chinese. This expansion facilitates cross-lingual evaluation of semantic contextualization and enables research into zero-shot transfer capabilities of multilingual models.

MCL-WiC (Martelli et al., 2021) offers datasets in English, Arabic, French, Russian, and Chinese. These were constructed by annotating sentences from native corpora, including BabelNet (Navigli and Ponzetto, 2012), the United Nations Parallel Corpus (Ziemski et al., 2016), and Wikipedia. The dataset achieved inter-annotator agreements of 0.95 and 0.9 for English and Russian, respectively, indicating high annotation quality.

AM^{2}iCo (Liu et al., 2021) presents a multilingual dataset pairing English with 14 target languages. Compiled from Wikipedia dumps of each

language, it selects words with at least two distinct pages, indicating ambiguity in both the target language and English. The dataset reports an overall human accuracy of 90.6% and an inter-annotator agreement of 88.4%.

WiC-TSV (Breit et al., 2021) introduces a multi-domain evaluation benchmark for WiC, independent of external sense inventories, but only in English. Covering various domains, WiC-TSV provides flexibility for evaluating diverse models and systems both within and across domains.

Despite these advancements, there remains a significant gap in resources for low-resource languages. Our dataset aims to address this deficiency by providing sense-annotated data in both WSD and WiC formats for underrepresented languages, thereby facilitating research in polysemy disambiguation and cross-lingual transfer across a broader spectrum of linguistic contexts.

3 Methods

3.1 Dataset Curation

We follow the below method for the curation of sense-annotated datasets, adjusted for language-specific considerations. These are detailed in section §4.1, along with the resources used for the curation of the dataset in each language.

1. Identification of Polysemous Words Publicly available dictionaries (online and offline) were surveyed. By searching for words with more than a single dictionary entry, lists of hundreds of candidate polysemous words were compiled. Where available, lists of polysemous words were added.

2. Corpus Selection and Sentence Sampling Native corpora of sufficient size were chosen to ensure diverse contextual representation of target words. Candidate polysemous words were filtered based on corpus frequency, removing low-frequency terms, and manually reviewed for sense granularity. From these corpora, large samples of sentences (typically 100-1000 per word) were randomly extracted for further analysis.

3. Embedding-Based Analysis Word embeddings were generated for target words in the sampled sentences, and dimensionality reduction methods and clustering techniques were applied to these to create interactive 2D visualizations (see §3.2).

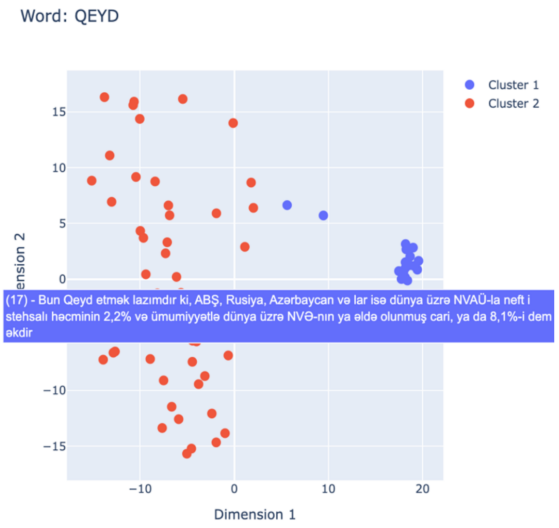


Figure 1: Example of interactive embedding-based sentence selection for the Azerbaijani word ‘qeyd’.

4. Manual Annotation of Sentences: In the 2D visualization, presented in Figure 1, annotators could hover over points representing sentences and click to assign them to different sense groups, for one word at a time. Sentences were selected based on their distribution in the embedding space or automatic clustering labels, with priority given to those that were more dispersed to ensure broad semantic coverage and enhance the representation of rare senses.

3.2 Semi-Automatic Annotation Tool

Our annotation process is semi-automatic, using vector representations for efficient sentence selection while ensuring manual verification.

To represent sentences in a structured way, we embed usages of the target word in all candidate sentences using pretrained transformer-based models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), or language-specific models. These embeddings capture contextual semantics, making them suitable for sense-based clustering. We then apply K-Means or agglomerative clustering to group sentences into distinct senses, followed by dimensionality reduction techniques (e.g., UMAP, MDS) to visualize their distribution in 2D space (see Figure 1).

This visualization allowed annotators to interact with embeddings, exploring clusters and selecting diverse sentences that represent different word senses. This is essential for identifying sentences that correspond to rare word senses, as manually

searching through randomly sampled sentences would be time-consuming and often ineffective, requiring the review of an extensive number of sentences to find relevant sentences.

3.3 Evaluating Annotation Efficiency

Annotating subordinate senses in polysemy is inherently time-consuming due to their rarity. Since these senses occur infrequently, manually identifying them requires scanning a large number of sentences before encountering a relevant instance.

The exact effort depends on the prior probability of the subordinate sense: the rarer it is, the more sentences need to be reviewed. To establish these priors, we randomly sampled 100 sentences for manual inspection to determine sense distributions. We then assessed how well model-based sentence selection captures each sense by comparing the proportion of automatically selected sentences correctly assigned to a sense against the baseline probability of encountering that sense in the corpus.

Our results demonstrate that computational methods significantly reduce this burden. We evaluate their effectiveness using adjusted **Lift**, a metric from Data Mining that measures improvement over random selection:

$$\text{Lift(sense)} = \frac{Precision(sense)}{Prior(sense)}$$

where *Precision(sense)* is the proportion of correctly classified sentences for the sense, and *Prior(sense)* their probability of occurrence in the dataset. Higher Lift values indicate a greater efficiency gain in selecting rare senses.

For example, in Kannada, identifying the word ಮತ in its subordinate 'religion' sense yielded a Lift of 900%, meaning that the likelihood of finding relevant sentences increased ninefold compared to random selection. Given a prior distribution of 96:4 favoring the dominant sense, manual selection would require reviewing 25 sentences on average to find one relevant case. With automatic selection achieving 36% precision, only three selections are needed—an 8× reduction in effort.

This efficiency boost translates directly into time and cost savings. If manual annotation takes 30 seconds per sentence, annotating 1,000 examples of a rare sense would traditionally require 8 hours of labor. With our automated method, this drops to about an hour, dramatically reducing annotation costs and making large-scale sense labeling more feasible. In Table 4, we present the Lift scores for the senses of two words in each of the four languages. Additional results, covering five words for each of these languages, are provided in Table B in the Appendix, covering all words selected for this evaluation.

Lang	Word	Sense Definitions		Lift (%)	
		1	2	1	2
KN	ಅಡಿ	Foot	Under	269	141
	ಮತ	Opinion	Religion	104	900
MR	रस	Juice	Interest	128	188
	उत्तर	Answer	North	121	125
PA	ਗੋਲੀ	Bullet	Pill	107	1364
	ਵਿਚਾਰ	Thought	Intention	235	884
UR	سونا	Gold	Sleep	161	232
	شکر	Thanks	Sugar	106	1414

Table 1: Measured improvement over random chance (Lift) in semi-automated sentence selection.

4 Sense-annotated Datasets

We introduce a sense-annotated corpus of sentences containing polysemous words covering ten low resource languages that span different language families and use different scripts: Azerbaijani (Turkic), Kannada and Telugu (Dravidian), Punjabi, Marathi and Urdu (Indo-Aryan), Polish (Slavic), Swahili (Afro-semitic), Vietnamese (Austroasiatic) and Korean (Koreanic). Statistics for each language are presented in Table 2.

4.1 Language Specific Treatment

For each language, the dataset was compiled and annotated by native speakers with the support of computational methods described above.

Azerbaijani: Polysemous words were selected from Azerbaycan Dilinin Omonimler Lugeti (Hesenov, 2007), and sentences containing selected target words were sampled from AzCorpus, the largest open-source NLP corpus for Azerbaijani (Kishiyev et al.). Three models were used to embed sentences: XLM-R, BERT-Turkish (DBMDZ, 2025), and XL-LEXEME (Cassotti et al., 2023).

Kannada: Polysemous words were selected from the online Kannada dictionary (Venkatasubbaiah et al., 1981). Kakwani et al. (2020a) was used as a corpus, which was preprocessed to remove extraneous characters, symbols, non-linguistic patterns, excessively long or single-word sentences, and duplicate entries. Initially, sentences for five words were annotated manually. Next, Claude 3.5 Sonnet (Anthropic, 2024) was used to pre-label

Language (ISO)	Words	Sentences	Senses	Avg. Senses/Word	Avg. Sentences/Sense
Azerbaijani (AZ)	60	4214	119	1.98 ± 0.13	35.55 ± 6.09
Kannada (KN)	59	4446	127	2.15 ± 0.45	35.01 ± 14.07
Korean (KO)	28	1013	58	2.07 ± 0.54	17.81 ± 4.73
Marathi (MR)	63	3766	125	1.98 ± 0.13	30.16 ± 2.72
Polish (PL)	66	2877	158	2.39 ± 0.68	18.28 ± 5.22
Punjabi (PA)	55	4969	127	2.31 ± 0.54	39.25 ± 1.89
Swahili (SW)	22	1376	46	2.09 ± 0.29	29.91 ± 4.39
Telugu (TE)	51	4534	100	1.96 ± 0.28	45.37 ± 7.83
Urdu (UR)	39	2674	90	2.31 ± 0.52	29.72 ± 1.06
Vietnamese (VI)	11	1021	29	2.64 ± 0.81	36.14 ± 19.20

Table 2: Statistics and ISO codes for the Multilingual WSD Sense-Annotated Dataset.

sentences after demonstrating reliable performance on the manually annotated data. The model, given Kannada and English meanings for each word, classified sentences containing the remaining target words. This streamlined human annotation, as annotators selected 30-40 sentences per sense from Claude's labels, rather than relying on clustering or embeddings for sentence selection. Finally, an independent reviewer verified all annotations.

Korean: The Korean Dictionary of National Institute of Korean Language (NIKL) (2025) was used to extract list of polysemous words. Two corpora were used for sampling sentences: the Korean Wikipedia Dataset (Lee, 2024) and KoWikiText (Kim, 2020). A Korean contextualized model (Ham et al., 2020) was used to embed sentences.

Marathi: The Marathi-English Dictionary from the Digital South Asia Library (DSAL) (Molesworth, 1857) was used to select polysemous words. For sampling sentences, three corpora were used: The Full Marathi Corpus (Joshi et al., 2022), and Marathi portions of two Indic corpora (Kakwani et al., 2020b; Kumar et al., 2023). MuRIL (Khanuja et al., 2021), mBERT, IndicBERT (Kakwani et al., 2020b), XLM-R, and XL-LEXEME were used for embedding sentences.

Polish: Polysemous words were identified by reviewing native texts, verified using the Polish Online Dictionary (Wydawnictwo Naukowe PWN, 2025), and selected if they had distinct senses. Three corpora covering distinct domains—national corpus, news, and literature—were used to sample sentences (Degórski and Przepiórkowski, 2012; Collection, 2018; Lebedev, 2023). XL-LEXEME and a Polish BERT (Kłeczek, 2020) were used for embedding sentences. Given Polish's high degree of inflection-where nouns, adjectives, and verbs vary by case, number, gender, and aspect across seven grammatical cases-all corpora were lemmatized to find sentences with target words in their base form for sentence selection and then restored to their original form for manual annotation.

Punjabi: Only text in Gurmukhi script was considered. Polysemous words were selected from previous work on WSD in Punjabi (Singh and Kumar, 2018, 2019, 2020; Singh and Singh, 2015) as well as from dictionaries (Joshi, 2009; Goswami, 2000; Brothers, 2006). Sentences were sampled from Metatext (Conneau et al., 2020), Samanantar (Ramesh et al.) and Sangraha (Khan et al.). MuRIL, IndicBERT, mBERT, XLM-R and XL-LEXEME were used to embed the sentences.

Swahili: The Swahili Dictionary (Chuo Kikuu cha Dar es Salaam, Taasisi ya Taaluma za Kiswahili, 2013) was used to identify polysemous words, while the Swahili Corpus by Masua and Masasi (2024) provided sentences. Multiple models were used for embedding (XLM-R, BERT and mBERT), but SwahBERT (Martin et al., 2022) outperformed them on the initial annotated dataset and was used to aid further annotation.

Telugu: Three corpora were used for selecting polysemous words, two Indic corpora (Kunchukuttan et al., 2020; Kakwani et al., 2020b) and the corresponding Wikipedia Dump (Wikimedia Foundation, 2024). The same Indic corpus (Kunchukuttan et al., 2020) was used for sentence selection, along with the Leipzig Telugu Corpus (Leipzig Corpora Collection, 2017). For embeddings, TeluguBERT (Joshi, 2022) and MuRIL were compared, with the former outperforming.

Urdu: Two word sense-annotated corpora (Saeed et al., 2019b,a), the Urdu Wiktextract (Ylonen, 2022), and a publicly available vocabulary

book (Bruce, 2021) were used to select polysemous words. The Urdu Monolingual Corpus (UrMono) (Jawaid et al., 2014) was used to sample sentences. For embedding, mBERT, XLM-R, MuRIL, and XL-LEXEME were tested with the latter outperforming the rest. Given Urdu's complex inflectional morphology and honorific system, a list of up to six inflected forms was generated for each noun, considering variations in number, gender, and case to ensure a diverse sentence selection.

Vietnamese: Polysemous words were selected from the Tuttle Concise Vietnamese Dictionary (Giuong, 2014), while sentences containing target words were sampled from the English-Vietnamese Parallel Corpus (EVBCorpus) (Ngo et al., 2013). For embedding, XL-LEXEME, XLM-R, mBERT, as well as two Vietnamese-specific models, PhoBERT (Nguyen and Tuan Nguyen, 2020) and ELECTRA (Nguyen, 2025) were evaluated. As with other languages, PhoBERT emerged as the best model, highlighting the need for language-specific methods and resources.

4.2 WiC Pairing

For model training we convert the sense-annotated data in each language to the WiC format (see §2.2).

To guarantee that the train-dev-test splits contain well-representative samples of words and sentences, and ensure sentences appear only in a single split, we use the following steps to convert sense-annotated sentences to WiC sentence pairs:

1. Word Splitting 70% of the words are randomly allocated to the training set, while 15% each are allocated to validation and test sets.

2. Sentence Redistribution 30% of words from the training set are randomly selected to appear in all three splits (each sentence appearing only in one of the splits). For these words, 25% of their sentences are reallocated to the validation and test sets, ensuring: (1) Equal distribution between sets; (2) No sentence overlap across splits; and (3) The distribution of senses remains unchanged.

3. Pairing Sentences into WiC Pairs Within each split, each sentence is paired with up to 16 different sentences, ensuring a balanced mix of same-sense and different-sense pairs.

The amounts were selected to approximate a 70-15-15 dataset split. This approach ensures a representative, well-distributed, and balanced dataset for WiC training and testing, although it's important to note that different random seeds for sampling can result in different results, especially for smaller datasets. Descriptive statistics of the resulting WiC datasets can be found in Table 5 in the Appendix. All sets are approximately balanced, setting chance performance close to 50%.

5 Experiments

To assess the quality of the datasets we created, and to demonstrate the need for proper evaluation in low-resource languages, we tested transfer in three transfer conditions, full-shot, zero-shot and mixed. The **full-shot** condition is mainly a sanity-check, and serves to evaluate the quality of the training set, as it does not test for transfer. In **zero-shot**, a model is fine-tuned on English (combined training data taken from the MCL (Martelli et al., 2021) and XL (Raganato et al., 2020) datasets, totaling 13.4k sentence pairs) and evaluated on each of our ten languages, which it was not fine-tuned on. In the **mixed** condition, a model is first fine-tuned on English, and then on the target language training data, evaluating on the target language. This allows us to investigate whether leveraging large amounts of data in a high-resource language can enhance full-shot performance on low-resource corpora.

We use XLM-RoBERTa (Conneau et al., 2020) due to its strong multilingual capabilities. The model is pretrained on 100 languages, including all those in our novel datasets. It has proven highly effective in embedding both high- and low-resource languages and is widely studied in cross-lingual transfer research (Philippy et al., 2023), particularly in the context of polysemy disambiguation (Raganato et al., 2020; Dairkee and Dubossarsky, 2024; Cassotti et al., 2023).

For model fine-tuning, we follow Cassotti et al. (2023) and use a bi-encoder architecture that independently processes the two sentences containing the polysemous target word using a Siamese network to generate two distinct vector representations (embeddings). The model outputs the cosine distance between the output embeddings of the two inputs, and, to collapse this to a binary label, a threshold is applied to decide if the words are classified as having the same sense. The model is trained to adapt embeddings and increase this distance when the target word has different meanings and decrease it when the meanings are the same in the two sentences by minimising contrastive loss. After training, we set the threshold for each model

Condition \ Test Lang	AZ	KN	KO	MR	PL	PA	SW	TE	UR	VI	Avg.
Full-shot	65.9	65.9	56.4	83.2	72.3	65.9	59.5	63.8	68.8	57.2	65.9
Zero-shot	66.3	**72.3**	64.2	82.2	**79.1**	70.5	68.6	62.4	**74.0**	**70.6**	71.0
Mixed	**71.9**	71.0	**66.5**	**88.1**	65.4	**81.6**	**76.9**	**65.4**	64.8	68.4	**72.0**

Table 3: Accuracies of XLM-R models evaluated on the test sets of our WiC datasets. Full-shot refers to models trained exclusively on the target language's training data. Zero-shot results correspond to XLM-R trained only on English WiC data. Mixed models are first trained on English, then fine-tuned on the target language.

by maximising accuracy on the corresponding validation set. During training, as well as inference, special tokens, <t> and </t>, are placed around the target word in each sentence to signal what word the model should focus on.

6 Results

Our semi-automatic annotation method works The transfer results (Table 3) demonstrate that we were able to produce high-quality datasets in ten low-resource languages. The low performance in Korean, Swahili, and Vietnamese is only observed in the full-shot condition. These are most likely due to their smaller training size rather than quality issues; otherwise, low performance would have been observed also in the zero-shot condition.

Evaluating on all target languages is essential Transfer effects are not uniform, as seen in the zero-shot performance that varies from 62.4% in Telugu to 82.2% in Marathi. Interestingly, zero-shot outperforms full-shot in 8 out of 10 languages, and gets comparable accuracy in the remaining 2, likely due to the small training data size of full-shot models and strong transfer from English. These results emphasize the unpredictability of transfer from one side, but also stress the need for a comprehensive multilingual benchmark to accurately assess cross-lingual transfer and ensure models perform reliably across diverse languages. With our efficient semi-automatic annotation method, curating such datasets is also much cheaper in annotation efforts.

Mixed training improves transfer For most languages, mixed-training improves upon either full-shot or zero-shot conditions. This hybrid strategy leverages large-scale training data in English with language-specific details from the target language for effective polysemy resolution. This further highlights the importance of datasets in low-resource languages, where even small amounts of labeled data can lead to marked improvements.

7 Discussion

In this work we present sense-annotated datasets across a diverse range of language families, providing valuable resources for linguistic and computational studies. Punjabi, Marathi, and Urdu belong to the Indo-Aryan branch, enabling research on linguistic relatedness alongside the Hindi WiC dataset (Dairkee and Dubossarsky, 2024). Telugu and Kannada represent the Dravidian family, while Azerbaijani, Swahili, Vietnamese, Polish, and Korean extend coverage to additional linguistic groups. The dataset includes Arabic-based (Punjabi, Urdu), Devanagari (Marathi), Latin-based (Azerbaijani, Polish, Swahili, Vietnamese), Hangul (Korean), and Brahmic scripts (Kannada, Telugu), facilitating research on script variation and its impact on NLP.

By encompassing a broad linguistic spectrum, our dataset supports studies on linguistic relatedness, historical evolution, and polysemy disambiguation in low-resource settings. It serves as a foundation for evaluating and improving multilingual and cross-lingual transfer, particularly in tasks requiring deep semantic understanding.

Our experiments highlight the importance of language-specific resources. The unexpected finding that zero-shot XLM-R trained only on English outperformed full-shot models trained on the target language challenges assumptions about cross-lingual transfer stability, emphasizing the need for dedicated evaluation datasets.

Manual annotation is essential yet labor-intensive, particularly for low-resource languages. We introduce an automated method to identify sentences across all word senses, even when certain senses are sparsely represented. Our quantitative results demonstrate the effectiveness of this approach in enhancing annotation efficiency and supporting sense-annotated dataset development. To encourage further research, we release our code on GitHub: github.com/roksanagow/projecting_sentences.

8 Limitations

The dataset remains relatively small, which may limit the generalizability of findings, particularly for full-shot experiments, where additional training data would likely improve performance. Additionally, data imbalance across languages makes direct comparisons challenging without subsampling, which in turn reduces overall performance. Even within a single language, the number of senses and sentences per word varies, further complicating evaluation. Moreover, each language was sourced from different corpora, leading to potential inconsistencies in text style, domain coverage, and annotation quality.

The evaluation setup also has certain constraints. Train-dev-test splits were generated randomly (according to the algorithm specified in §4.2), and the prevalence of sentences corresponding to different words across splits could impact the results. Furthermore, zero-shot evaluation was conducted only from English, leaving open questions about transfer from other high-resource languages and cross-lingual settings beyond English-centric transfer.

9 Acknowledgments

This work was partially funded by the research program Change is Key!, supported by Riksbankens Jubileumsfond (reference number M21-0021). The authors would like to thank Bao Linh Hoang for contributing additional expert annotations.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Google AI. 2018. Multilingual bert (mbert). Accessed: August 2024.

Anthropic. 2024. Claude 3.5 sonnet. Available at https://www.anthropic.com/news/claude-3-5-sonnet.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*, pages 4330–4338. International Joint Conference on Artificial Intelligence, Inc.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2554–2560. Association for Computational Linguistics.

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for target sense verification of words in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645. Association for Computational Linguistics.

Singh Brothers. 2006. *Punjabi-English Dictionary*. Singh Brothers, Amritsar. Accessed: August 2024.

Gregory Maxwell Bruce. 2021. *Urdu Vocabulary: A Workbook for Intermediate and Advanced Students*. Edinburgh University Press, Edinburgh. Accessed: August 2024.

Agostina Calabrese, Michele Bevilacqua, Roberto Navigli, et al. 2020. Fatality killed the cat or: Babelpic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4680–4686. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708, Tokyo, Japan. Association for Computational Linguistics.

Chuo Kikuu cha Dar es Salaam, Taasisi ya Taaluma za Kiswahili. 2013. *Kamusi ya Kiswahili Sanifu*. Oxford University Press, East Africa Limited, Nairobi, Kenya. Accessed: August 2024.

Leipzig Corpora Collection. 2018. pol_news_2018: Polish news corpus. https://corpora.wortschatz-leipzig.de/ic?corpusId=pol_news_2018. Accessed: August 2024.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Farheen Dairkee and Haim Dubossarsky. 2024. Strengthening the wic: New polysemy dataset in hindi and lack of cross lingual transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15341–15349.

DBMDZ. 2025. BERT-Base Turkish 128K Uncased. Available at Hugging Face, Accessed: August 2024.

Łukasz Degórski and Adam Przepiórkowski. 2012. Recznie znakowany milionowy podkorpus nkjp. *Przepiórkowski et al.[17]*, pages 51–58.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805.

Błażej Dolicki and Gerasimos Spanakis. 2021. Analysing the impact of linguistic features on cross-lingual transfer. *arXiv preprint arXiv:2105.05975*.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: Overview. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5. Association for Computational Linguistics.

Phan Van Giuong. 2014. *Tuttle Concise Vietnamese Dictionary: Vietnamese-English, English-Vietnamese*. Tuttle Publishing, North Clarendon, VT. Accessed: August 2024.

K. K. Goswami. 2000. *Punjabi-English/English-Punjabi Dictionary*. Hippocrene Books, New York. Accessed: August 2024.

J. Ham, Y. J. Choe, K. Park, I. Choi, and H. Soh. 2020. Ko-sroberta multitask model. Available at Hugging Face, Accessed: August 2024.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Hesret Hesenov. 2007. Azerbaycan dilinin omonimler lugeti. *Serq Qerb nesriyyati. Baki*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Bushra Jawaid, Amir Kamran, and Ondřej Bojar. 2014. A tagged corpus and a tagger for urdu. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2938–2943, Reykjavik, Iceland. European Language Resources Association (ELRA). Accessed: August 2024.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*. https://arxiv.org/abs/2211.11418.

Raviraj Joshi et al. 2022. L3Cube-MahaNLP: Marathi natural language processing datasets, models, and library. Accessed: August 2024.

S. S. Joshi. 2009. *Punjabi-English Dictionary: Panjabi Yuniwarasiti Panjabi-Angarezi Kosha*. Punjabi University, Patiala. Accessed: August 2024.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020a. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020b. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Accessed: August 2024.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad G, Varun Balan G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Hyunjoong Kim. 2020. Kowikitext: A wikitext format korean corpus. Accessed: August 2024.

Huseyn Kishiyev, Jafar Isbarov, Kanan Suleymanli, Khazar Heydarli, Leyla Eminova, and Nijat Zeynalov. azcorpus: The largest open-source nlp corpus for azerbaijani (1.9m documents, 18m sentences). Accessed: August 2024.

Anoop Kumar et al. 2023. Sangraha: A high-quality, multilingual dataset for indic language pretraining. Accessed: August 2024.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. *arXiv preprint arXiv:2005.00085*. Accessed: August 2024.

Darek Kłeczek. 2020. Polbert: Polish bert language model. Accessed: August 2024.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Dmitrii Lebedev. 2023. Polish classic literature text corpus. Accessed: August 2024.

Chang W. Lee. 2024. Wikipedia korean dataset (2024-05-01). Accessed: August 2024.

Leipzig Corpora Collection. 2017. Telugu community corpus (2017). Accessed: August 2024.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Federico Martelli, Najla Kalach, Gabriele Tola, Roberto Navigli, et al. 2021. Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (mcl-wic). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36.

Gati Martin, Medard Edmund Mswahili, Young-Seob Jeong, and Jiyoung Woo. 2022. SwahBERT: Language model of Swahili. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–313, Seattle, United States. Association for Computational Linguistics.

Bernard Masua and Noel Masasi. 2024. In the heart of swahili: An exploration of data collection methods and corpus curation for natural language processing. *Data in Brief*, 55:110751.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

James Thomas Molesworth. 1857. *A Dictionary, Marathi and English*, 2 edition. Printed for government at the Bombay Education Society's Press, Bombay. Accessed: August 2024.

National Institute of Korean Language (NIKL). 2025. Nikl korean-english dictionary. Dataset available on Hugging Face. Accessed: August 2024.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli et al. 2023. BabelNet 2023. `https://babelnet.org/publications`.

Quoc Hung Ngo, Werner Winiwarter, and Bartholomäus Wloka. 2013. EVBCorpus - a multi-layer english-vietnamese bilingual corpus for studying tasks in comparative linguistics. In *Proceedings of the 11th Workshop on Asian Language Resources (ALR-11) at IJCNLP 2013*, pages 1–9, Nagoya, Japan. Asian Federation of Natural Language Processing. Accessed: August 2024.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Nha Van Nguyen. 2025. Nlphust/ner-vietnamese-electra-base. Accessed: August 2024.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019a. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019b. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293, Brussels, Belgium. Association for Computational Linguistics.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. 10:145–162.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):13.

Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2019a. A sense annotated corpus for all-words urdu word sense disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 18(4):1–14. Accessed: August 2024.

Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2019b. A word sense disambiguation corpus for urdu. *Language Resources and Evaluation*, 53(3):397–418. Accessed: August 2024.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Jagbir Singh and Iqbal Singh. 2015. Word sense disambiguation: Enhanced lesk approach in punjabi language. *International Journal of Computer Applications*, 129(6):23–27. Accessed: August 2024.

Varinder Pal Singh and Parteek Kumar. 2018. Naive bayes classifier for word sense disambiguation of punjabi language. *Malaysian Journal of Computer Science*, 31(3):188–199. Accessed: August 2024.

Varinder Pal Singh and Parteek Kumar. 2019. Sense disambiguation for punjabi language using supervised machine learning techniques. *Sādhanā*, 44(11):226. Accessed: August 2024.

Varinder Pal Singh and Parteek Kumar. 2020. Word sense disambiguation for punjabi language using deep learning techniques. *Neural Computing and Applications*, 32(8):2963–2973. Accessed: August 2024.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual nlp models. *arXiv preprint arXiv:2110.08875*.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

G. Venkatasubbaiah, L. S. Sheshagiri Rao, and H. K. Ramachandra. 1981. *Kannada-Kannada-English Dictionary*. Ibh Prakashana, Bangalore, India. Accessed: August 2024.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wikimedia Foundation. 2024. Telugu wikipedia dump. Accessed: August 2024.

Wikimedia Foundation. 2025. List of wikipedias. Accessed: August 2024.

Wydawnictwo Naukowe PWN. 2025. Słownik języka polskiego pwn. Accessed: August 2024.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France. European Language Resources Association. Accessed: August 2024.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Example visualization of annotated sentences

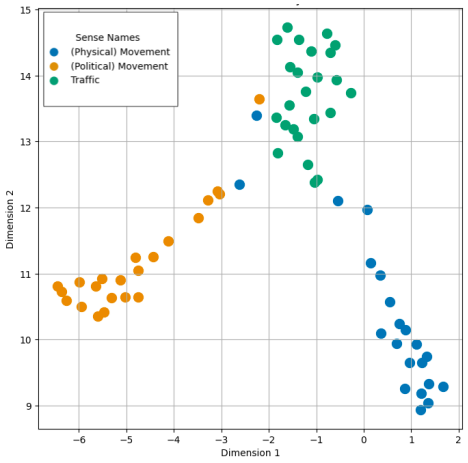


Figure 2: Word embeddings of the Polish word 'ruch' in sense-annotated sentences, visualized in 2D with UMAP. Interestingly, the resulting shape resembles a walking figure.

B Evaluating Annotation Efficiency

Lang	Word	Sense Definitions			Lift (%)		
		1	2	3	1	2	3
KN	ಅಡಿ	Foot	Under	-	269	141	-
	ಮತ	Opinion	Religion	-	104	900	-
	ಗುಂಡಿ	Pit/Hole	Bullet	-	269	141	-
	ಮಂಡಿ	Market	Knee	-	167	223	-
	ಮಾತು	Word	Conversation	-	289	128	-
MR	रस	Juice	Interest	-	228	288	-
	उत्तर	Answer	North	-	221	225	-
	मान	Respect	Approval	-	218	235	-
	खोली	Room	Depth	-	147	124	-
	हार	Necklace	Defeat	-	106	149	-
PA	ਗੋਲੀ	Bullet	Pill	-	107	1364	-
	ਵਿਚਾਰ	Thought	Intention	-	235	884	-
	ਉੱਤਰ	North	Response	Descend	210	438	156
	ਖਾਨ	Khan (name)	Mine	-	128	211	-
	ਹਾਰ	Defeat	Necklace	-	129	∞ (prior = 0)	-
UR	سونا	Gold	Sleep	-	161	232	-
	شکر	Thanks	Sugar	-	106	1414	-
	زبان	Language	Tongue	-	119	358	-
	کانٹا	Thorn	Fork	-	108	808	-
	اتفاق	Opportunity	Agreement	Coincidence	685	155	364

Table 4: Measured improvement over random chance (Lift) in semi-automated sentence selection over all evaluated words.

C WiC sentence pairing

Language	AZ	KN	KO	MR	PL	PA	SW	TE	UR	VI
Sent Pairs (Train)	20,409	20,298	5,703	19,368	13,516	26,237	7,312	23,115	14,018	5,153
Sent Pairs (Dev)	5,649	5,627	1,018	5,175	3,562	7,025	2,165	5,861	3,450	751
Sent Pairs (Test)	5,434	4,809	656	4,194	3,103	5,749	1,100	5,500	3,210	1,397
Words (Train)	42	42	20	45	47	39	16	36	28	8
Words (Dev)	22	22	11	24	25	21	9	19	15	5
Words (Test)	22	21	9	22	24	19	7	18	14	4
Words in All Splits	13	13	6	14	15	12	5	11	9	3

Table 5: Amounts of sentence pairs and unique polysemous target words in the train-dev-test splits of our constructed WiC datasets.

XCOMPS: A Multilingual Benchmark of Conceptual Minimal Pairs

Linyang He[*1] **Ercong Nie**[*2,3]
Sukru Samet Dindar[1] **Arsalan Firoozi**[1] **Adrian Florea**[1]
Van Nguyen[3] **Corentin Puffay**[5] **Riki Shimizu**[1] **Haotian Ye**[2,3]
Jonathan Brennan[4] **Helmut Schmid**[3] **Hinrich Schütze**[2,3†] **Nima Mesgarani**[1†]
[1]Columbia University [2]Munich Center for Machine Learning
[3]LMU Munich [4]University of Michigan [5]KU Leuven
linyang.he@columbia.edu nie@cis.lmu.de
hinrich@hotmail.com nima@ee.columbia.edu

Abstract

In this work, we introduce XCOMPS, a multilingual conceptual minimal pair dataset that covers 17 languages. Using this dataset, we evaluate LLMs' multilingual conceptual understanding through metalinguistic prompting, direct probability measurement, and neurolinguistic probing. We find that: 1) LLMs exhibit weaker conceptual understanding for low-resource languages, and accuracy varies across languages despite being tested on the same concept sets. 2) LLMs excel at distinguishing concept-property pairs that are visibly different but exhibit a marked performance drop when negative pairs share subtle semantic similarities. 3) More morphologically complex languages yield lower concept understanding scores and require deeper layers for conceptual reasoning. The dataset is publicly available at: https://github.com/LinyangHe/XCOMPS/.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across various natural language understanding (NLU) tasks. Recent advances, such as GPT-4 (Achiam et al., 2023) and Llama 3 (Dubey et al., 2024), have shown that LLMs can produce human-like outputs and handle complex linguistic phenomena. However, whether LLMs genuinely understand semantics or merely rely on shallow statistical correlations is disputable (Lake and Baroni, 2018; Elazar et al., 2021; Huang et al., 2023). One fundamental aspect of human conceptual understanding is that it is not dependent on specific linguistic forms or modalities (Carey, 2000; Mandler, 2004). When humans learn and reason about concepts, they do not require the knowledge to be tied to a particular medium, such as text, images, or video, nor do they rely on a specific language. This raises an important question: *Does LLMs' conceptual-property reasoning remain stable across languages, or is it language-specific?*

To explore this, Misra et al. (2023) introduced the COMPS dataset, designed to probe the semantic reasoning abilities of LLMs through minimal pairs in English. However, COMPS only evaluates monolingual conceptual-property reasoning, leaving open the question of whether LLMs generalize such reasoning across languages. In this work, we introduce XCOMPS, a multilingual extension of COMPS, to assess whether LLMs' semantic reasoning is universally consistent across languages. XCOMPS covers 17 languages, including analytic, inflectional, and agglutinative languages, ensuring a broad representation of linguistic structures.

Beyond dataset expansion, evaluating LLMs' reasoning abilities has increasingly relied on prompt engineering, often referred to as metalinguistic prompting (Hu and Levy, 2023). However, recent work (Hu and Levy, 2023; He et al., 2024b) suggests that metalinguistic prompting primarily assesses performance—that is, how well a model produces correct outputs—rather than its underlying competence in conceptual understanding. This distinction is crucial, as models may perform well on explicit prompts but lack true conceptual representations (Piantadosi and Hill, 2022). To investigate LLMs' multilingual capabilities and determine whether they genuinely encode conceptual knowledge across languages, we adopt a three-pronged evaluation approach: *Metalinguistic prompting*, *Neurolinguistic probing*, and *Direct probability measurement*. Our experimental results reveal several insights into the multilingual conceptual reasoning capabilities of LLMs: 1) Conceptual understanding is not consistently maintained across languages. Even when models perform well in English, their reasoning ability deteriorates significantly in low-resource languages; the extent of deterioration also varies across different low-resource languages. 2) Models perform well when concep-

[*] Equal contribution.
[†] Corresponding authors.

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 75–81
August 1, 2025 ©2025 Association for Computational Linguistics

Type	Language	Acceptable Sentence	Unacceptable Sentence
Taxonomic	Spanish	*Tostadora* se utiliza para calentar alimentos. (A toaster is used for heating food.)	*Cafetera* se utiliza para calentar alimentos. (A coffee maker is used for heating food.)
Overlap	Vietnamese	*Máy nướng bánh mì được* sử dụng để hâm nóng thực phẩm. (A toaster is used for heating food.)	*Tủ lạnh được* sử dụng để hâm nóng thực phẩm. (A refrigerator is used for heating food.)
Co-occurrence	Hungarian	*Kenyérpirító* ételek melegítésére használják. (A toaster is used for heating food.)	*Vízforraló* ételek melegítésére használják. (A kettle is used for heating food.)
Random	Dutch	*Broodrooster* wordt gebruikt om voedsel te verwarmen. (A toaster is used for heating food.)	*Winterkoning* wordt gebruikt om voedsel te verwarmen. (A wren is used for heating food.)

Table 1: XCOMPS examples, illustrating each linguistic variant pairs an acceptable sentence (positively matched property) with an unacceptable counterpart (negatively matched property).

tual relationships are highly distinct but struggle with subtle semantic distinctions. 3) Languages with higher morphological complexity (agglutinative > inflected > analytic) yield lower concept-reasoning scores. These results suggest that LLMs' semantic reasoning may not generalize universally across linguistic boundaries.

2 Language Performance vs. Competence

As suggested in He et al. (2024b), LLMs can be evaluated through three methods: *metalinguistic prompting*, which assesses *performance* based on explicit responses; direct probability measurement, which provides an intermediate evaluation by comparing model-generated probabilities; and *neurolinguistic probing*, which directly examines *competence* by analyzing internal activation patterns[1].

Metalinguistic Prompting for Performance This method involves explicitly querying the model about linguistic expressions, often in a comparative or multiple-choice format. By asking the model to choose between minimal pairs (e.g., "Which sentence is more grammatically correct?"), researchers can evaluate how well the model retrieves and verbalizes knowledge. Using prompting, researchers have revealed new classes of emergent abilities such as arithmetic, instruction-following, grounded conceptual mappings, and sentence acceptability judgments (Brown et al., 2020; Wei et al., 2022; Patel and Pavlick, 2021; Dentella et al., 2023). Because the responses are influenced by prompt engineering and surface-level cues, this method primarily reflects performance rather than deep conceptual competence.

Direct Probability Measurement Instead of relying on explicit responses, this method examines the model's probability assignment to different sentences within minimal pairs. For example, a model should assign a higher probability to 'A robin can fly' than to 'A penguin can fly'. This approach offers a more objective evaluation than metalinguistic prompting and captures implicit model preferences, placing it between performance and competence. Researchers have designed syntactic, semantic/conceptual, and discourse inference tasks using the probability assignment method, offering different insights into LLMs' capabilities compared to metalinguistic prompting (Futrell et al., 2019; Gauthier et al., 2020; Hu et al., 2020; Warstadt et al., 2020; Beyer et al., 2021; Misra et al., 2023; Kauf et al., 2023). However, it still relies on external outputs and does not fully reveal how the model internally represents concepts.

Neurolinguistic Probing for Competence This approach goes beyond external outputs by analyzing internal activation patterns across different layers of the model (He et al., 2024a,b). Using diagnostic classifiers, researchers can probe whether LLMs inherently encode conceptual-property relationships or simply rely on statistical correlations. Since it provides a direct measure of competence, neurolinguistic probing is more reliable for assessing the depth of linguistic understanding.

3 XCOMPS

3.1 Concept Selection

To ensure that XCOMPS maintains conceptual alignment with COMPS while extending its scope to multiple languages, we use the same 521 concepts and their negative samples from COMPS. As shown in Table 1, these negative samples can be categorized into three types. *Taxonomy-based* negative samples are selected based on hierarchical relationships among concepts. Negative samples come from the same broad category as the positive concept but differ in key property attributions. *Property norm-based (overlap)* negative samples are chosen based on shared semantic properties with the positive concept while lacking the specific

[1]For simplicity, we refer to these three methods as Meta, Direct, Neuro.

property under evaluation. *Co-occurrence-based samples* are selected from concepts that frequently appear in similar contexts but do not share the target property. XCOMPS also has additional *random negative concepts* from the set of concepts that do not possess the property of the original positive concept.

3.2 Properties of Concepts

In XCOMPS, the properties assigned to concepts are inherited from COMPS, ensuring alignment across languages while maintaining the original conceptual-property relationships. These properties in COMPS were originally derived from the XCSLB dataset, an extended version of the CSLB property norm dataset (Devereux et al., 2014), which captures human-annotated perceptual, functional, and categorical attributes of concepts. Additionally, taxonomic relationships from resources like WordNet (Miller, 1995) were used to infer properties through hierarchical inheritance, ensuring that general category attributes (e.g., "mammals have fur") are systematically applied to their subcategories. Some properties also reflect real-world associations observed in corpus-based co-occurrence statistics.

3.3 Multilingual Data Construction

To construct XCOMPS, which covers 17 languages (Table 2 in Appendix A), we adopted a human-LLM interactive translation pipeline, leveraging both human expertise and the multilingual generation capabilities of LLMs. The language set for XCOMPS aligns with the prior knowledge probing benchmarks, such as BMLAMA-17 (Qi et al., 2023) and KLAR (Wang et al., 2025), ensuring consistency in multilingual evaluation. The highly structured nature of conceptual minimal pair datasets, where positive and negative sentences primarily consist of two components–concepts and properties–enabled us to design a multi-step translation process that ensures high-quality multilingual data.

The construction process consists of four stages. We use the GPT-4o model (GPT-4o-2024-08-06) via the OpenAI API as the translation assistant in the pipeline. In the first stage, we manually translated the original concepts and properties from English into German and Chinese using language experts. We used German and Chinese as additional seed languages to further reduce ambiguity, This multilingual seed data helped disambiguate concepts that might otherwise be unclear in translation. For example, the English word "bat" could refer to either the flying animal or the sports equipment. By including the German term "Schläger" and the Chinese term "球拍", which both unambiguously refer to the sports equipment, we ensured that the intended concept was accurately captured during translation.

In the second stage, we used LLMs to expand the seed data into the remaining 15 languages. LLMs were tasked with translating the concepts and properties, leveraging their multilingual machine translation capabilities. By providing seed data in three languages (English, German, and Chinese), we enhanced the LLMs' ability to generate accurate translations, as the additional context reduced the likelihood of semantic errors.

In the third stage, human experts for each target language manually reviewed and corrected the translated concepts and properties. This step ensured that the translations were accurate, culturally appropriate, and semantically aligned with the original dataset. Human intervention was particularly critical for low-resource languages, where LLMs often struggle with semantic precision in translation tasks.

Finally, in the fourth stage, LLMs were employed to generate complete sentences based on the verified concepts and properties. This step involved formulating positive and negative sentence pairs, which can be viewed as a straightforward language manipulation task. By providing the translated concepts and properties as input, we enabled the LLMs to focus on generating fluent and grammatically correct sentences, leveraging their strengths in multilingual text generation. This approach ensured that the most challenging aspect of the task–accurate translation of concepts and properties–was already resolved, allowing the LLMs to produce high-quality outputs.

By splitting the process into property translation and sentence generation, using multilingual seed data to reduce ambiguity, and combining human expertise with LLM capabilities, we ensured the quality and consistency of the XCOMPS dataset. This human-LLM interactive translation pipeline demonstrates how LLMs' multilingual understanding and generation capabilities can be effectively harnessed to construct high-quality multilingual benchmarks.

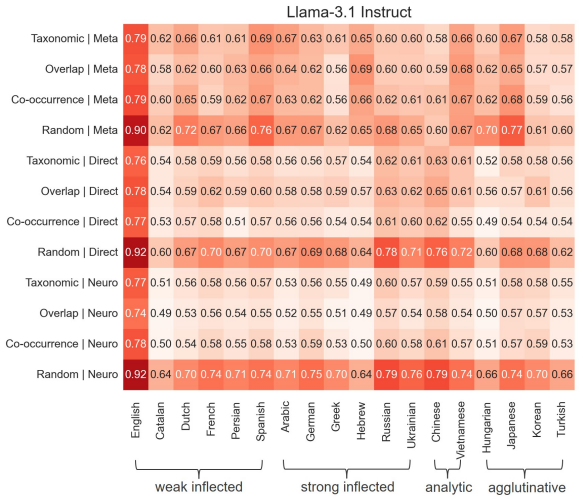


Figure 1: Metalinguistic prompting (meta), direct probability measurement (direct), and minimal pair probing (neuro) results on XCOMPS. The meta method evaluates LLMs' language performance; the neuro method evaluates LLMs' language competence, and the direct method falls in between. Languages are grouped according to morphological typology. Neuro-probing is a layer-wise method, and here we use the max value across all layers to compare with Meta and Direct.

4 Experiment Setup

4.1 Model

We use meta-llama/Llama-3.1-8B-Instruct from Hugging Face in our experiment, which applies instruction tuning to the base model for more intuitive user-prompt handling. During the inference, we adopt float16 precision to minimize computational resource consumption while maintaining model performance.

4.2 Evaluation

For **Meta**, we present both sentences of a minimal pair within a single prompt. We convert the target property into a question and compare the probabilities assigned to acceptable vs. unacceptable concepts. Figure 2 in Appendix A shows the prompts used in the experiment. For **Direct**, we compute sentence probabilities directly from the model's logits. A prediction is considered correct if the model assigns a higher probability to the valid sentence within each minimal pair. For **Neuro**, we adopt last-token pooling to represent each sentence, extracting the final token's hidden state from every layer. This approach ensures coverage of all preceding tokens (Meng et al., 2024). We then apply a logistic regression classifier for probing, using the F1 score (averaged over five cross-validation folds) as our primary evaluation metric.

4.3 Results and Analysis

Cross-linguistic variability in conceptual reasoning. From Figure 1, we observe that the model can perform relatively well on English conceptual tasks but show marked declines for low-resource languages. Notably, some languages with limited training data (e.g., Hungarian, Catalan) exhibit greater deterioration than others, indicating that cross-linguistic generalization of conceptual understanding is far from uniform. Even within the low-resource category, the degree of performance drop varies, underscoring that LLMs' semantic reasoning is neither universally stable nor equally supported by existing multilingual corpora. These patterns reinforce the idea that conceptual capabilities learned in English do not necessarily transfer seamlessly to languages that differ typologically or have weaker representations in training data.

Models excel at distinct conceptual contrasts but falter with subtler differences. High scores all appear in Random rows, where the negative concept is clearly distinct (e.g., "toaster" vs. "wren"), and the model easily detects mismatches. In Taxonomic, Overlap, or Co-occurrence rows, however, performance drops because the negative concepts share subtle semantic similarities (e.g., "toaster" vs. "coffee maker"). This indicates that the models may rely on conspicuous cues rather than true conceptual reasoning.

Direct and neuro convergence. By comparing direct and neuro results in Figure 1, and from Figure 3 in Appendix A, we see high correlations across all negative types, indicating that direct measurements closely track the models' internal representations.

Higher morphological complexity, lower conceptual reasoning. Figure 4 in Appendix A shows that languages with greater morphological complexity (moving from Analytic to Inflected to Agglutinative) tend to yield lower concept-reasoning scores. This indicates that, as linguistic structure becomes more complex, it becomes harder for the models to capture concept-property relationships consistently.

5 Conclusion

In this work, we introduce the XCOMPS benchmark, which provides a multilingual conceptual

minimal pair dataset for evaluating the language model's semantic understanding across 17 languages. This work reveals that while LLMs demonstrate surface-level multilingual capabilities, they lack a universal semantic reasoning mechanism that transcends language boundaries.

Limitation

While XCOMPS significantly advances the evaluation of multilingual conceptual understanding, certain limitations remain. First, although the dataset covers 17 typologically diverse languages, it does not encompass all linguistic families or low-resource languages, which may limit its generalizability to underrepresented languages. Second, the reliance on human-LLM interaction for data construction ensures high quality but introduces potential inconsistencies due to variations in human expertise and model outputs. Lastly, while XCOMPS focuses on conceptual understanding, it does not explicitly address other challenges in multilingual NLP, such as pragmatics or contextual reasoning. Despite these limitations, XCOMPS provides a robust foundation for assessing and improving LLMs' multilingual capabilities, and future work can extend its scope to address these areas.

Acknowledgement

We thank the anonymous reviewers for their valuable advice and feedback. This research was partially supported by DFG (German Research Foundation) grant SCHU 2246/14-1 and Munich Center for Machine Learning (MCML).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. *arXiv preprint arXiv:2105.03495*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melvin Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33.

Susan Carey. 2000. The origin of concepts. *Journal of Cognition and Development*, 1(1):37–41.

Vittoria Dentella, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2023. Testing ai performance on less frequent aspects of language reveals insensitivity to underlying meaning. *arXiv preprint arXiv:2302.12313*.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46:1119–1127.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R Brennan. 2024a. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. 2024b. Large language models as neurolinguistic subjects: Identifying internal representations for form and meaning. *arXiv preprint arXiv:2411.07533*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen,

Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Jean Matter Mandler. 2004. *The foundations of mind: Origins of conceptual thought*. Oxford University Press.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2920–2941.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*.

Steven T Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. *arXiv preprint arXiv:2504.04264*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

A Appendix

Table 2 shows the detailed information of the languages covered by XCMOPS. Figure 2 displays the prompt templates of different languages used for metalinguistic prompting evaluation. Figures 3 and 4 show detailed experimental results.

lid	language	Typology	Family
ar	Arabic	Inflectional	Semitic
ca	Catalan	Inflectional	Indo-European (Romance)
de	German	Inflectional	Indo-European (Germanic)
el	Greek	Inflectional	Indo-European (Hellenic)
es	Spanish	Inflectional	Indo-European (Romance)
fa	Persian	Inflectional	Indo-European (Iranian)
fr	French	Inflectional	Indo-European (Romance)
he	Hebrew	Inflectional	Semitic
hu	Hungarian	Agglutinative	Uralic
ja	Japanese	Agglutinative	Isolate
ko	Korean	Agglutinative	Isolate
nl	Dutch	Inflectional	Indo-European (Germanic)
ru	Russian	Inflectional	Indo-European (Slavic)
tr	Turkish	Agglutinative	Turkic
uk	Ukrainian	Inflectional	Indo-European (Slavic)
vi	Vietnamese	Analytic	Austroasiatic
zh	Chinese	Analytic	Sino-Tibetan

Table 2: Detailed information of the languages covered by XCOMPS.

```
en: Which concept is most likely to have the following property: "{property}", "{word1}" or "{word2}"? Answer: "
ar: أي مفهوم من المرجح أن يكون لديه الخاصية التالية: "{property}"، "{word1}" أو "{word2}"؟ الإجابة: "
he: איזה מושג סביר ביותר שיש לו את המאפיין הבא: "{property}", "{word1}" או "{word2}"? תשובה: "
fa: کدام مفهوم به احتمال زیاد دارای ویژگی زیر است: "{property}"، "{word1}" یا "{word2}"؟ پاسخ: "
de: Welches Konzept hat am wahrscheinlichsten die folgende Eigenschaft: "{property}", "{word1}" oder "{word2}"? Antwort: "
zh: 哪个概念最有可能有如下特征: "{property}", "{word1}" 还是 "{word2}"? 回答: "
fr: Quel concept est le plus susceptible d'avoir la propriété suivante : "{property}", "{word1}" ou "{word2}" ? Réponse : "
nl: Welk concept heeft waarschijnlijk de volgende eigenschap: "{property}", "{word1}" of "{word2}"? Antwoord: "
es: ¿Qué concepto es más probable que tenga la siguiente propiedad: "{property}", "{word1}" o "{word2}"? Respuesta: "
ja: どの概念が次の特性を持つ可能性が最も高いですか: "{property}"、"{word1}" または "{word2}"? 答え: "
ko: 어떤 개념이 다음 속성을 가질 가능성이 가장 높습니까: "{property}", "{word1}" 또는 "{word2}"? 답변: "
vi: Khái niệm nào có khả năng nhất có thuộc tính sau: "{property}", "{word1}" hoặc "{word2}"? Câu trả lời: "
el: Ποια έννοια είναι πιο πιθανό να έχει την ακόλουθη ιδιότητα: "{property}", "{word1}" ή "{word2}"? Απάντηση: "
hu: Melyik fogalomnak van a legnagyobb esélye, hogy rendelkezik a következő tulajdonsággal: "{property}", "{word1}" vagy "{word2}"? Válasz: "
tr: Hangi kavramın şu özelliğe sahip olma olasılığı daha yüksektir: "{property}", "{word1}" veya "{word2}"? Cevap: "
ca: Quin concepte és més probable que tingui la següent propietat: "{property}", "{word1}" o "{word2}"? Resposta: "
uk: Яке поняття найімовірніше має таку властивість: "{property}", "{word1}" чи "{word2}"? Відповідь: "
ru: Какое понятие с наибольшей вероятностью обладает следующим свойством: "{property}", "{word1}" или "{word2}"? Ответ: "
```

Figure 2: Prompt templates of different languages used for metalinguistic prompting.

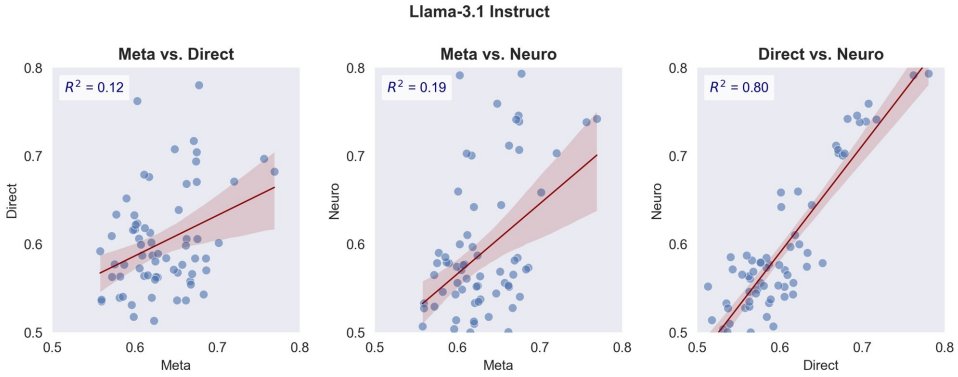


Figure 3: Linear correlation among meta, direct, and neuro evaluation results for all four tasks.

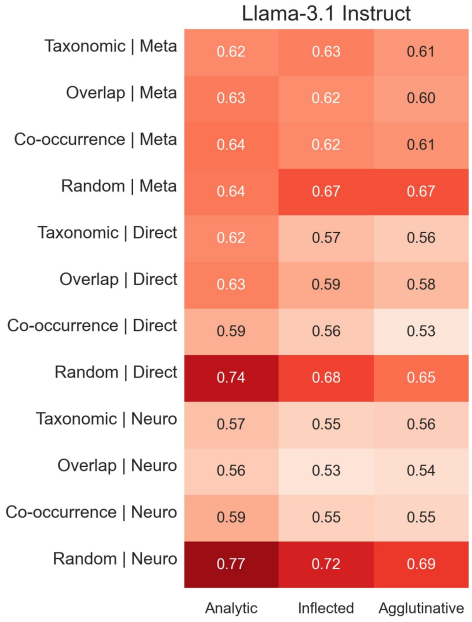


Figure 4: Averaged results across different language types. English results are dropped to make the comparison more reliable among low-resource languages.

Tone in Perspective: A Computational Typological Analysis of Tone Function in ASR

Siyu Liang and **Gina-Anne Levow**
University of Washington
liangsy, levow@uw.edu

Abstract

This study investigates the impact of pitch flattening on automatic speech recognition (ASR) performance across tonal and non-tonal languages. Using vocoder-based signal processing techniques, we created pitch-flattened versions of speech recordings and compared ASR performance against original recordings. Results reveal that tonal languages experience substantially larger performance degradation than non-tonal languages. Analysis of tone confusion matrices shows systematic patterns of misidentification where contour tones collapse toward level tones when pitch information is removed. Calculation of tone's functional load at syllable and word levels demonstrates that syllable-level functional load strongly predicts ASR vulnerability to pitch flattening, while word-level patterns reflect each language's morphological structure. These findings illuminate the differential importance of pitch information across languages and suggest that ASR systems for languages with high syllable-level functional load require more robust pitch modeling.

1 Introduction

Lexical tone, where pitch distinctions signal differences in word meaning, is a core feature of over half the world's languages (Yip, 2002). While tonal contrasts rely primarily on fundamental frequency (f_0), they also interact with duration, intensity, and voice quality. These complexities pose unique challenges for automatic speech recognition (ASR), particularly in tonal languages where pitch plays a central role in lexical identity.

Recent ASR models implicitly encode tonal information, but it remains unclear how critical pitch actually is for recognition across language types. To investigate this, we apply pitch flattening—a signal processing technique that removes f_0 contours—to speech recordings and compare ASR performance with and without flattened pitch contours across both tonal (Thai, Vietnamese, Mandarin) and non-tonal (Uzbek, Indonesian, Turkish) languages.

We find that tonal languages experience significantly larger degradation in ASR performance under pitch flattening, with systematic tone confusion patterns revealing that contour tones (e.g., falling, rising) tend to collapse toward level tones when f_0 contours are removed. To explain these differences, we compute the functional load of tone and show that syllable-level functional load is a strong predictor of ASR vulnerability, capturing cross-linguistic differences in tone dependency more effectively than word-level metrics.

2 Background and Related Work

2.1 Tone

Tone refers to the use of pitch patterns to distinguish lexical or grammatical meanings, and it appears in over half of the world's languages (Yip, 2002). At its core, tone is related to fundamental frequency (f_0), often supplemented by secondary cues such as duration or phonation type (e.g., creaky or breathy voice) (Garellek et al., 2013; Zhang and Kirby, 2020). While languages like Thai, Vietnamese, and Mandarin all employ pitch contrasts, each does so differently: Thai traditionally has five tones, Vietnamese features six, and Mandarin typically has four plus a neutral tone (Yip, 2002; Thurgood, 2002). The functional importance of tone also varies cross-linguistically; in some systems, pitch shapes nearly every syllable, whereas others use additional cues for lexical contrasts.

From a linguistic perspective, these pitch contrasts often evolve through *tonogenesis*—the historical development of tone from segmental distinctions such as voicing (Haudricourt, 1954). Once established, tone can become as critical as vowels or consonants in signaling word meaning (Surendran and Levow, 2004). This high informational

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 82–92
August 1, 2025 ©2025 Association for Computational Linguistics

load means that even small shifts in f_0 may yield major changes in lexical interpretation. Yet tone is not always "standalone": interactions with intonation, stress, or morphology can influence its role within the broader phonological system.

2.2 Tone and ASR

The significance of pitch in tone perception poses unique challenges for ASR technology. Early systems for Chinese and Thai explicitly modeled pitch tracks alongside spectral features (Fu et al., 1998; Lei et al., 2006), while modern end-to-end frameworks often rely on learned representations (e.g., XLS-R (Babu et al., 2021)) to capture tonal nuances. Even so, how effectively these systems handle pitch remains an open question—particularly for low-resource tonal languages, where sparse training data compound recognition errors (Coto-Solano, 2021; Qin et al., 2022).

2.3 Pitch Manipulation

One way to isolate pitch's contribution is *pitch flattening*, which systematically removes f_0 contours while preserving segmental and temporal information (Valbret et al., 1992). This technique has informed both psycholinguistic studies—showing how listeners rely on other cues like duration or context when pitch is lost (Wang et al., 2013)—and ASR research, where drops in recognition accuracy can reveal a system's reliance on pitch. Related work has compared natural speech against flattened or synthesized stimuli for languages such as Mandarin and Thai (Liu and Samuel, 2004; Zsiga and Nitisaroj, 2007), demonstrating substantial performance declines in human perception when f_0 cues are removed or distorted.

2.4 Functional Load

To quantify how critical pitch distinctions are in any given language, researchers often invoke *functional load* (Hockett, 1967; Surendran and Levow, 2004). This information-theoretic metric captures the extent to which a contrast (e.g., a particular tone versus no tone) contributes to lexical distinctions. Languages with a high tonal load—where a substantial portion of the semantic space hinges on pitch—are predictably more vulnerable when pitch cues degrade. In contrast, languages whose words can be distinguished by segmental or morphological features may be less affected by pitch flattening.

2.5 Tone and Typology

Because tone systems vary dramatically, from heavily monosyllabic languages like Vietnamese to those where multisyllabic words dilute the burden on pitch (Thurgood, 2002; Brunelle and Kirby, 2016), cross-linguistic experimentation is pivotal for robust ASR design. Studies have shown that, in some languages, phonation features may help compensate for reduced f_0 (Brunelle and Kirby, 2016), while in others, listeners (and ASR systems) default to level or "unmarked" tones when pitch is unavailable (Francis et al., 2003). By comparing both tonal and non-tonal languages under pitch-flattened conditions, we can pinpoint how different phonological structures handle the loss of f_0 cues and where ASR systems might fail. Insights from such comparisons suggest which modeling strategies, e.g., explicit pitch tracking, tone-based lexicons, or phonation-sensitive acoustic features, offer the most gains for languages heavily reliant on pitch.

3 Methods

We designed experiments to evaluate how pitch manipulation influences ASR performance across typologically diverse languages. Specifically, we investigate how removing lexical pitch cues via pitch flattening affects recognition accuracy in tonal versus non-tonal languages. By comparing ASR performance on original and pitch-flattened versions of the same utterances, we aim to quantify the importance of pitch information for recognition and identify the linguistic and structural factors that predict vulnerability to pitch manipulation.

3.1 Data

We selected six languages for our study: three tonal languages (Thai, Vietnamese, and Mandarin Chinese) and three non-tonal languages (Uzbek, Indonesian, and Turkish). Our selection of tonal languages was primarily constrained by data availability in the speech corpora and is typologically biased toward East and Southeast Asian tone systems. While these languages represent important tone types, they do not capture the full typological diversity of tone systems found worldwide, such as register tone languages of Africa or pitch-accent systems, which will be discussed in Section 7. All data were drawn from the Common Voice 17.0 corpus (Ardila et al., 2020). For each language, we used 2 hours of speech data for training and 30

Language	Original Text	Processed Text
Thai	ผมรักเธอ	phoom4 rak1 thoe0
Vietnamese	Tôi yêu bạn	tôi1 yêu1 ban6
Mandarin	我爱你	wo3 ai4 ni3

Table 1: Text preprocessing examples for “I love you” in the three tonal languages, showing original text and preprocessed text.

minutes for testing. All audio data were resampled at 16 kHz.

3.2 Preprocessing

For non-tonal languages, we applied minimal processing (standardized case and removed punctuation). For tonal languages, we applied specific preprocessing to ensure consistent transcription for tones. Table 1 shows examples of this preprocessing for each language.

For Thai, we used pythainlp.transliterate with engine=tltk_g2p, which converts Thai script to Latin characters with explicit tone marking (numbers 0–4). The numeric tone markers correspond to: 0 = mid tone, 1 = low tone, 2 = falling tone, 3 = high tone, and 4 = rising tone. Note that tone numbers used here follow a phonological convention rather than pitch height, where, for example, *rak1* (“love”) is a mid-tone syllable (not high), resulting from a low-class consonant with a dead syllable and no tone mark. In Vietnamese, we mapped diacritics denoting tone to numeric tone labels while keeping other diacritics for vowel contrast intact. Our mapping converted Vietnamese diacritics to numeric tone labels as follows: 1 = ngang (level/no diacritic), 2 = huyền (falling/grave accent), 3 = sắc (rising/acute accent), 4 = hỏi (dipping/hook), 5 = ngã (creaky/tilde), and 6 = nặng (heavy/dot below). For Mandarin Chinese, we used the pypinyin package with style=Style.TONE3. The numeric markers correspond to: 1 = high level tone (āi), 2 = rising tone (ái), 3 = falling-rising tone (ǎi), 4 = falling tone (ài), without explicitly including the neutral tone.

3.3 Pitch Flattening

Pitch flattening was performed using Praat’s Pitch-Synchronous OverLap and Add (PSOLA) algorithm (Valbret et al., 1992). This procedure effectively neutralizes lexical tone cues while maintaining other speech properties, including duration, intensity, and spectral envelope. In our implementation, the f_0 contour of each utterance was replaced with the utterance’s mean f_0 value. Figure 1 illustrates the process on a sample Thai utterance, showing the original and flattened pitch contours.

We should note that flattening the contour does not eliminate every trace of pitch, as micro-periodicity cues remain in the harmonic spectrum. Therefore, our results are a conservative estimate of tone dependence; a future experiment that additionally uses the interharmonic energy of low-pass filters would provide an even “cleaner” ablation.

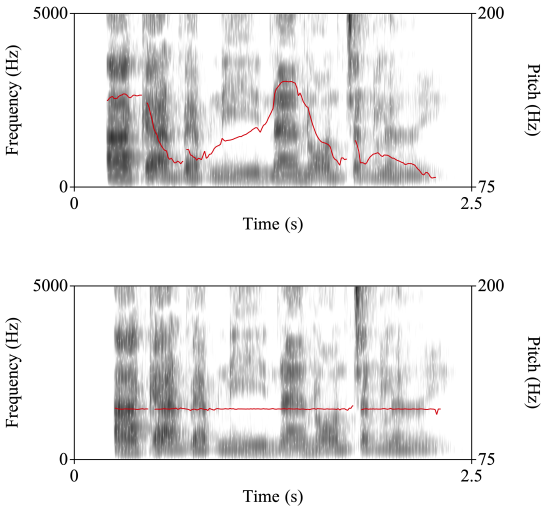


Figure 1: Example of pitch flattening on a Thai utterance "This kind of weather makes me feel sleepy." The top panel shows the original spectrogram overlaid with pitch contour; the bottom panel shows the flattened version of the same audio.

3.4 ASR Model Training

We fine-tuned individual XLS-R 300m models (Babu et al., 2021) for each language. Specifically, we trained the model on 2 hours of speech from Common Voice 17.0 and tested on 30 minutes. Additionally, for each tonal language, we ran the ASR model on pitch flattened test data too. Hyperparameters and training details are included in the Appendix (see Appendix A.1 for complete hyperparameter settings).

3.5 Evaluation Metrics

We evaluated ASR performance using multiple metrics to capture different aspects of recognition accuracy. In addition to WER (Word Error Rate) and CER (Character Error Rate), we also use additional metrics given in Table 2.

Metric	Description
TER	Tone Error Rate: percentage of syllables with correctly recognized segments but incorrectly identified tones
ConER	Consonant Error Rate: errors in consonant recognition
VER	Vowel Error Rate: errors in vowel recognition
WER-T	Modified version of WER where tone markers were ignored
CER-T	Modified version of CER where tone markers were ignored
Δ	Absolute difference between pitch-flattened and original speech for each metric

Table 2: Evaluation metrics used to assess ASR performance across languages.

3.6 Tone Confusion Analysis

For tonal languages, we constructed tone confusion matrices to analyze specific patterns of tone misidentification when pitch information was removed. These matrices recorded the counts of each reference tone (true label) being recognized as each possible tone (predicted label) in both original and pitch-flattened conditions. We then calculated difference matrices (flattened minus original) to identify which tonal confusions increased most dramatically after pitch flattening.

3.7 Functional Load Calculation

To quantify the information-theoretic contribution of tone in each language, we calculated the functional load (FL) of tonal contrasts at both syllable and word levels, following the methodology of Surendran and Levow (2004):

$$FL = \frac{H_{with} - H_{without}}{H_{with}} \quad (1)$$

where H_{with} represents the Shannon entropy of the distribution with tonal contrasts maintained, and $H_{without}$ represents the entropy after neutralizing tonal distinctions.

For syllable-level calculations, we extracted syllable frequencies from our corpus, maintaining or neutralizing tone distinctions to compute the respective entropies. For word-level calculations, we employed language-specific tokenization tools: PyThaiNLP with the newmm engine for Thai, Jieba for Mandarin, and underthesea for Vietnamese. These tools provided morphological segmentation used for analyzing the relationship between tone and word structure.

We also calculated the average number of syllables per word for each language to understand how morphological characteristics might influence the relationship between syllable-level and word-level functional loads. These calculations allowed us to quantitatively assess whether languages with higher functional load of tone would show greater vulnerability to pitch flattening in ASR performance.

4 Results

4.1 Impact of Pitch Flattening on ASR Performance

Table 3 presents our baseline ASR outcomes for six languages (three tonal, three non-tonal), comparing recognition on the original recordings vs. pitch-flattened audio that removes f_0 contours. As expected, the tonal languages (Vietnamese, Mandarin, Thai) experience substantially larger performance drops than the non-tonal ones (Uzbek, Indonesian, Turkish), confirming that pitch serves as a crucial contrastive cue for tone-based systems.

In particular, Thai displays the highest jump in WER upon flattening (+0.232), with Mandarin and Vietnamese also incurring significant degradations (+0.194 and +0.118). By contrast, pitch removal in Uzbek, Indonesian, and Turkish increases WER by only 5–8 points, indicating that segmental cues alone largely suffice for lexical discrimination in these atonal settings.

4.2 Tone Dependence and Detailed Phonetic Metrics

To examine tone-dependence in further detail, Table 4 shows additional metrics for the three tonal languages, including *tone error rate* (TER), *consonant error rate* (ConER), *vowel error rate* (VER), and error rates when ignoring tone markers (WER-T, CER-T). Thai exhibits the largest TER increase ($+0.2543$), reflecting its strong reliance on f_0 cues. Mandarin and Vietnamese also display pronounced TER jumps of +0.2009 and +0.1837, respectively.

Although consonant and vowel error rates increase less dramatically, they still reveal that pitch flattening affects the broader phonetic structure, not only the tonal dimension. When ignoring tone, i.e., disregarding tone output in error rate calculation, the error rates CER-T and WER-T of the three tonal languages are very similar to the non-tonal languages in Table 3.

4.3 Tone Confusion

Figure 2 illustrates the changes in tone confusion patterns after pitch flattening. More details about the values can be found in Appendix A.2. Each

Language	WER (orig.)	WER (flat.)	Δ_{WER}	CER (orig.)	CER (flat.)	Δ_{CER}
Tonal						
Vietnamese	0.715	0.833	0.118	0.312	0.380	0.068
Mandarin	0.478	0.672	0.194	0.209	0.283	0.074
Thai	0.288	0.520	0.232	0.082	0.154	0.072
Non-Tonal						
Uzbek	0.782	0.857	0.075	0.247	0.288	0.041
Indonesian	0.599	0.668	0.069	0.193	0.232	0.039
Turkish	0.743	0.816	0.073	0.240	0.292	0.052

Table 3: WER and CER results under original vs. pitch-flattened conditions, grouped by tonal and non-tonal categories. The Δ columns show $(\text{Flattened - Original})$.

Language	Version	TER	Δ_{TER}	ConER	Δ_{ConER}	VER	Δ_{VER}	WER-T	$\Delta_{\text{WER-T}}$	CER-T	$\Delta_{\text{CER-T}}$
Vietnamese	original	0.3954		0.3525		0.3739		0.6430		0.3063	
	flattened	0.5791	0.1837	0.3929	0.0404	0.4199	0.0460	0.6932	0.0502	0.3408	0.0345
Mandarin	original	0.3430		0.4300		0.3287		0.6169		0.4646	
	flattened	0.5439	0.2009	0.4658	0.0358	0.3686	0.0399	0.6838	0.0669	0.5066	0.0420
Thai	original	0.1266		0.0981		0.0864		0.2465		0.0810	
	flattened	0.3809	0.2543	0.1279	0.0298	0.1205	0.0341	0.3099	0.0634	0.1087	0.0277

Table 4: Comparison of tone error rate (TER), consonant error rate (ConER), vowel error rate (VER), and ignoring-tone WER/CER for Vietnamese, Mandarin, and Thai.

heatmap plots the *difference* (flattened minus original counts), where red regions indicate increased confusion and blue regions show decreased confusion. Analysis of these patterns reveals specific directional shifts in tone recognition after f_0 removal.

Across all three languages, diagonal elements (representing correct tone identification) show substantial negative values, indicating significantly reduced accuracy. Thai exhibits the largest average diagonal decrease (-146.40 per tone), followed by Mandarin (-232.25) and Vietnamese (-94.50). Conversely, off-diagonal elements show positive values (Thai: +29.28, Vietnamese: +15.36, Mandarin: +56.75), reflecting increased confusion between different tones.

The most pronounced confusion patterns are highly directional. In Thai, flattened audio led to falling tone being misidentified as mid tone (+246 instances), followed by rising tone confused with mid tone (+111). This suggests that without f_0 contours, the distinctive falling and rising patterns collapse toward the perceptually less marked mid tone. Thai's falling tone showed the largest proportional decrease in correct identification (-55.2%), followed by rising tone (-43.1%).

Vietnamese exhibited a striking trend where multiple tones were confused with ngang (level) tone after flattening: huyền (falling) $\rightarrow$ ngang (+312), sắc (rising) $\rightarrow$ ngang (+259), hỏi (dipping) $\rightarrow$ ngang (+92), and nặng (heavy) $\rightarrow$ ngang (+56). This systematic shift toward the unmarked ngang tone demonstrates how pitch flattening neutralizes the distinctive contour features of Vietnamese tones. The huyền tone showed the most dramatic reduction in correct identification (-46.9%), while the ngang tone was least affected.

For Mandarin, the most significant confusion was falling tone misidentified as high tone (+306), followed by rising tone confused with high tone (+129). Without pitch cues, distinctive contour tones (falling, rising, fall-rise) are increasingly confused with the level high tone. The falling tone experienced the largest decrease in accuracy (-30.6%), consistent with its heavily pitch-dependent contour.

These directional confusions reveal a general pattern: in the absence of f_0 contrast, contour tones (those with dynamic pitch movements such as falling, rising, or complex contours) collapse toward level tones (mid tone in Thai, ngang in Vietnamese, and high tone in Mandarin). While the results are consistent with the idea that level tones function as unmarked defaults, they could equally reflect an artefact of the acoustic manipulation: the loss of dynamic contour cues renders rising, falling, and dipping tones indistinguishable. We caution, however, that flattened utterances are acoustically

Language	Syllable FL	Word FL	Avg. Syll./Word	Δ_{WER}	Δ_{TER}
Thai	0.1243	0.0189	1.86	0.232	0.2543
Mandarin	0.0597	0.0336	1.15	0.194	0.2009
Vietnamese	0.0530	0.0517	0.99	0.118	0.1837

Table 5: Functional load (FL) of tone at syllable and word levels, with average syllables per word and ASR performance degradation metrics.

atypical for any training distribution. Some of the observed errors may thus reflect domain mismatch rather than pure loss of lexical information.

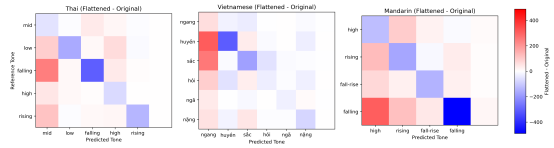


Figure 2: Confusion matrices based on tone count difference comparing flattened to original audio for Thai, Vietnamese, and Mandarin. Red cell marks increased prediction in that category, and blue cell marks decreases. Off-diagonal hotspots reveal a consistent drift of contour tones toward each language's level tone (mid, ngang, and high, respectively) in the left column.

4.4 Functional Load and Tone Dependency

To better understand the relationship between tone importance and ASR degradation, we calculated the functional load (FL) of tone at both syllable and word levels across the three tonal languages based on 32k tokens from the transcripts of the same Common Voice database, with the data scarcity of Vietnamese as the lower bound. Table 5 summarizs the results and reveals an interesting pattern: syllable-level functional load aligns remarkably well with ASR performance degradation. Thai exhibits the highest syllable-level FL (0.1243), followed by Mandarin (0.0597) and Vietnamese (0.0530), a ranking that precisely mirrors the order of WER increase under pitch flattening (Thai: +0.232, Mandarin: +0.194, Vietnamese: +0.118) and TER increase (Thai: +0.2543, Mandarin: +0.2009, Vietnamese: +0.1837). This strong correlation (Pearson's $r = 0.91$ for syllable FL vs. WER degradation) suggests that syllable-level functional load effectively predicts a language's ASR vulnerability to pitch flattening.

Interestingly, word-level functional load presents a different pattern. Vietnamese maintains nearly all of its tonal information at the word level (word FL: 0.0517, 97.5% of its syllable FL), while Mandarin preserves about half (word FL: 0.0336, 56.3% of syllable FL), and Thai retains only 15.2% (word FL: 0.0189). These proportions directly reflect each language's morphological structure: Vietnamese's predominantly monosyllabic words (average 0.99 syllables per word) necessitate tone distinctions for lexical identity, whereas Thai's higher proportion of multisyllabic words (average 1.86 syllables per word) allows tone to function more as one feature among many for word identification.

This morphological analysis complements our earlier confusion matrix findings. In Vietnamese, where tone information remains critical at the word level, confusion patterns show tones collapsing toward the less marked ngang (level) tone, but overall ASR degradation is less severe than in languages with higher syllable-level functional load. Thai, despite maintaining less tone information at the word level, experiences the largest performance drop precisely because its syllable-level tone distinctions carry substantial information that cannot be compensated for by other features when pitch is removed.

The pattern of flattening-induced confusion (contour tones collapsing toward level tones) observed in Figure 2 offers additional insight into why languages with higher syllable-level functional load suffer greater ASR degradation. Languages where tone carries more syllable-level information typically employ more distinctive contour tones, which are particularly vulnerable to pitch flattening. This vulnerability manifests in the dramatic decreases in recognition accuracy for falling (-55.2%) and rising (-43.1%) tones in Thai.

Taken together, these findings suggest that syllable-level functional load offers a more effective predictor of ASR vulnerability to pitch degradation than word-level measures. This has important implications for speech technology development across tonal languages: systems for languages with high syllable-level functional load will require more robust pitch modeling and may benefit from explicit tone-specific accommodations, while those for languages with lower tone dependency might be more resilient to noisy pitch environments.

5 Discussion

Our results reveal significant differences in how the ASR results of tonal and non-tonal languages respond to pitch flattening, with systematic patterns that illuminate the relationship between tone, speech perception, and ASR performance. These findings have important implications for both lin-

guistic theory and speech technology development.

5.1 Differential Impact of Pitch Flattening

The substantially larger ASR performance degradation observed in tonal languages (Thai: +23.2%, Mandarin: +19.4%, Vietnamese: +11.8% WER) compared to non-tonal languages (5-8% WER increase) confirms the critical role of f_0 information in tonal language processing. However, the non-zero impact on non-tonal languages indicates that pitch also contributes to speech recognition even when not lexically contrastive, likely through prosodic cues that help segment and identify words.

The varying degrees of degradation among tonal languages suggest differences in tone dependency. Thai showed the highest vulnerability to pitch flattening. This could be explained by our functional load analysis revealed Thai has a higher syllable-level tonal information density. These results align with Surendran and Levow (2004), who found language-specific differences in tone's functional load, but extend their work by demonstrating a direct relationship between this information-theoretic measure and ASR vulnerability.

The relatively smaller impact on Vietnamese (+11.8% WER) despite its complex six-tone system suggests that Vietnamese ASR benefits from additional disambiguating cues. As noted by Brunelle and Kirby (2016), Vietnamese tones involve substantial phonation contrasts (creaky, breathy voice) that may provide redundant information when pitch cues are removed. This phonation-based redundancy appears to partially compensate for the loss of f_0 information in Vietnamese, unlike in Thai and Mandarin where pitch plays a more singular role.

5.2 Tone Confusion Patterns and Perceptual Structure

The tone confusion analysis revealed striking directional patterns across all three tonal languages. In Thai, falling and rising tones were frequently confused with mid tone; in Vietnamese, multiple tones collapsed toward ngang (level) tone; and in Mandarin, contour tones were often misidentified as high tone. This systematic shift of confusion from contour tones toward level tones suggests that with neutralized f_0 cues, ASR systems default to perceptually unmarked tonal categories, a finding that parallels observations in human speech perception studies (Francis et al., 2003; Khouw and Ciocca, 2007).

It should be noted that pitch-flattened syllables are not strictly equivalent to natural 'level tones' in these languages. Natural level tones in East and Southeast Asian languages also include pitch movements, such as a slight fall or rise at the end, and are produced with specific phonation characteristics (Yip, 2002). Despite this distinction, our results show that when pitch information is neutralized through flattening, ASR systems consistently default to categorizing these flattened stimuli as level tones, suggesting that level tones serve as defaults in the absence of distinctive pitch movement.

These directional confusions have both acoustic and phonological implications. Acoustically, contour tones (with dynamic pitch movements) are more dependent on f_0 information than level tones. Phonologically, the patterns align with markedness theory: level tones typically function as unmarked categories in tonal systems (Yip, 2002), serving as defaults when distinctive features are unavailable. Importantly, this pattern is not simply a frequency effect, such as evident in our Mandarin data (see Appendix A.2) where the falling tone (4) is actually the most frequent in our dataset, yet confusion still predominantly shifts toward the high level tone (1) rather than following raw frequency distributions.

The diagonal values in the confusion matrices (representing correct identification) showed the largest decreases for tones with substantial pitch movement: falling tone in Thai (-55.2%), huyền tone in Vietnamese (-46.9%), and falling tone in Mandarin (-30.6%). This suggests that the perceptual distance between tones is not uniform but depends on their phonetic realization, with contour tones being perceptually more distant from other categories and thus more vulnerable to pitch flattening.

5.3 Functional Load and Language Structure

Our functional load analysis provides a quantitative framework for understanding cross-linguistic differences in tone dependency. The strong correlation between syllable-level functional load and ASR degradation (Thai: 0.1243/+23.2% WER, Mandarin: 0.0597/+19.4% WER, Vietnamese: 0.0530/+11.8% WER) suggests that this information-theoretic measure effectively predicts a language's vulnerability to pitch flattening.

The differences between syllable-level and word-level functional load reflect each language's morphological structure. Vietnamese maintained nearly all its tonal information at the word level (97.5%

of syllable-level FL), consistent with its predominantly monosyllabic nature. By contrast, Thai preserved only 15.2% of its tonal information at the word level, reflecting its higher proportion of multisyllabic words where tone distinctions on individual syllables become less critical for overall word identification.

These patterns highlight an important insight: a language's dependency on tone is not solely determined by the number of tonal contrasts or their acoustic properties, but also by the information-theoretic role of tone within the broader phonological and morphological system. Languages with high syllable-level functional load, especially those with significant proportions of monosyllabic words, are inherently more vulnerable to pitch perturbations.

5.4 Implications for ASR Development

Our findings have several practical implications for ASR system development. First, they suggest that pitch modeling requirements differ substantially across languages, even among those classified as tonal. Languages with high syllable-level functional load (like Thai) would benefit from explicit modeling of pitch contours, while those with redundant cues (like Vietnamese) might achieve acceptable performance with less sophisticated pitch representations.

Second, the systematic tone confusion patterns identified could inform error correction strategies in ASR systems. By understanding the likely confusion directions when pitch information is degraded (e.g., contour tones being misidentified as level tones), post-processing algorithms could apply targeted corrections based on contextual and acoustic cues.

Third, our results suggest that ASR robustness for tonal languages could be improved through explicit modeling of phonation cues, particularly for languages like Vietnamese where voice quality provides redundant information. Integrating both pitch and phonation features would create systems more resilient to acoustic degradations affecting either dimension.

Fourth, language modeling capabilities could potentially compensate for degraded tonal information. Our experiments used a basic CTC-based approach without additional language modeling, but we hypothesize that stronger language models could help recover tone information from context in pitch-degraded scenarios. This could be particularly effective in languages with higher word-level redundancy, where contextual cues might disambiguate tonally similar syllables.

Finally, the functional load framework offers a principled approach for predicting a priori which languages will require more sophisticated tone modeling in ASR systems. Rather than treating all tonal languages uniformly, developers could allocate resources based on information-theoretic measures of tone's importance in each language.

6 Conclusion

This study investigated the impact of pitch flattening on ASR performance across tonal and non-tonal languages, revealing several key insights about the role of pitch in speech recognition. Our findings demonstrate that tonal languages experience substantially greater performance degradation when pitch information is removed, but with significant variations that correlate with the functional load of tone in each language. The systematic patterns of tone confusion observed—where contour tones collapse toward level tones—highlight fundamental aspects of tonal perceptual structure.

Beyond documenting these effects, we established a quantitative relationship between information-theoretic measures of tone importance and ASR vulnerability. Languages with high syllable-level functional load proved most susceptible to pitch flattening, while word-level functional load patterns reflected each language's morphological characteristics. This framework offers a principled approach for predicting which languages will require more sophisticated tone modeling in speech technology applications.

Our findings have implications for both linguistic theory and ASR system development. Theoretically, they support models of tone perception where unmarked level tones serve as default categories when distinctive pitch information is unavailable. Practically, they suggest that ASR systems for tonal languages should be designed with language-specific considerations of tone's functional load and the availability of redundant acoustic cues.

Future work could extend this analysis to a wider typological range of tone systems. For instance, examining Cantonese, which features a more complex inventory of level tones, could test whether our observed pattern of confusion toward level tones holds in languages where multiple level tones must

be distinguished. Similarly, investigating Bantu languages, which feature tonal contrasts that are often analyzed differently from East Asian systems despite having contour properties, would broaden our typological understanding of how different tone systems respond to pitch degradation.

7 Limitations

While providing valuable insights, our study has several limitations that suggest directions for future research. First, our analysis focused on ASR performance rather than human perception. Parallel studies with human listeners would clarify whether the confusion patterns observed are specific to machine learning systems or reflect broader perceptual principles.

Second, our pitch flattening approach, while effective at isolating the contribution of f_0, represents an extreme case of pitch degradation. Future work could explore more nuanced manipulations, such as partial flattening or targeted disruption of specific pitch features, to identify which aspects of the pitch contour are most critical for recognition.

Third, our functional load calculations were limited to tone's contribution and did not address interactions with other phonological features. Expanding this analysis to include phonation, vowel quality, and other features would provide a more comprehensive understanding of how different dimensions contribute to lexical contrasts across languages.

Fourth, our ASR system used basic CTC-based decoding without sophisticated language modeling. A stronger language model would likely improve overall performance and might partially compensate for pitch flattening through contextual prediction. Future work should investigate the degree to which language modeling can mitigate the effects of degraded tonal information in various languages.

Finally, while we included three major tonal languages, our study does not capture the full typological diversity of tone systems. Extending this work to include languages with different tonal inventories (e.g., Cantonese with its multiple level tones), register tone languages (e.g., Hmong), pitch-accent languages (e.g., Japanese), and languages with different tone systems like those found in Bantu languages would provide a more complete picture of how pitch information contributes to speech recognition across language types.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. *arXiv preprint*. ArXiv:1912.06670 [cs].

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint*. ArXiv:2111.09296 [cs, eess].

Marc Brunelle and James Kirby. 2016. Tone and Phonation in Southeast Asian Languages: Tone and Phonation in Southeast Asian Languages. *Language and Linguistics Compass*, 10(4):191–207.

Rolando Coto-Solano. 2021. Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.

Alexander L. Francis, Valter Ciocca, and Brenda Kei Chit Ng. 2003. On the (non)categorical perception of lexical tones. *Perception & Psychophysics*, 65(7):1029–1044.

Qian-Jie Fu, Fan-Gang Zeng, Robert V Shannon, and Sigfrid D Soli. 1998. Importance of tonal envelope cues in Chinese speech recognition. *The Journal of the Acoustical Society of America*, 104(1):505–510. Publisher: Acoustical Society of America.

Marc Garellek, Patricia Keating, Christina M. Esposito, and Jody Kreiman. 2013. Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America*, 133(2):1078–1089.

André-Georges Haudricourt. 1954. De l'origine des tons en vietnamien. *Journal Asiatique*, 242:69–82.

Charles F. Hockett. 1967. The Quantification of Functional Load. *WORD*, 23(1-3):300–320. Publisher: Routledge _eprint: https://doi.org/10.1080/00437956.1967.11435484.

Edward Khouw and Valter Ciocca. 2007. Perceptual correlates of Cantonese tones. *Journal of Phonetics*, 35(1):104–117.

Xin Lei, Manhung Siu, Mei-Yuh Hwang, Mari Ostendorf, and Tan Lee. 2006. Improved tone modeling for Mandarin broadcast news speech recognition. In *Interspeech 2006*, pages paper 1752–Tue3A2O.4–0. ISCA.

Siyun Liu and Arthur G. Samuel. 2004. Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech*, 47(Pt 2):109–138.

Siqing Qin, Longbiao Wang, Sheng Li, Jianwu Dang, and Lixin Pan. 2022. Improving low-resource Tibetan end-to-end ASR by multilingual and multilevel unit modeling. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):2.

Dinoj Surendran and Gina-Anne Levow. 2004. The functional load of tone in Mandarin is as high as that of vowels. In *Speech Prosody 2004*, pages 99–102. ISCA.

Graham Thurgood. 2002. Vietnamese and tonogenesis: Revising the model and the analysis. *Diachronica*, 19(2):333–363.

H. Valbret, E. Moulines, and J. P. Tubach. 1992. Voice transformation using PSOLA technique. *Speech Communication*, 11(2):175–187.

Jiuju Wang, Hua Shu, Linjun Zhang, Zhaoxing Liu, and Yang Zhang. 2013. The roles of fundamental frequency contours and sentence context in Mandarin Chinese speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1):EL91–EL97.

Moira Jean Winsland Yip. 2002. *Tone*. Cambridge textbooks in linguistics. Cambridge University Press, Cambridge ; New York.

Yubin Zhang and James Kirby. 2020. The role of F0 and phonation cues in Cantonese low tone perception. *The Journal of the Acoustical Society of America*, 148(1):EL40–EL45.

Elizabeth Zsiga and Rattima Nitisaroj. 2007. Tone Features, Tone Perception, and Peak Alignment in Thai. *Language and Speech*, 50(3):343–383. Publisher: SAGE Publications Ltd.

A Appendix

This appendix provides additional details on our fine-tuning hyperparameters for XLS-R 300m in both experiments.

A.1 XLS-R Fine-Tuning Hyperparameters

All training runs (for both Common Voice and TIBMD@MUC data) used the same set of essential hyperparameters, with only minor adjustments for batch size depending on GPU memory:

- **Model:** facebook/wav2vec2-xls-r-300m
- **Batch Size:** 8
- **Learning Rate:** 3×10^{-4}
- **Warmup Steps:** 500
- **Max Steps:** 2000
- **Vocabulary Size:** based on unique characters in the training corpus (including space or | as word delimiter).

A.2 Tone Confusion Results

The results of tone confusion are as follows:

	0	1	2	3	4	none
0	912	25	25	33	13	3
1	36	473	17	16	15	1
2	35	10	496	14	3	0
3	35	12	15	287	3	0
4	12	23	5	12	274	1
none	0	0	0	0	0	0

Table 6: Thai tone confusion (**Original**). Rows = reference tone (0 = Mid, 1 = Low, 2 = Falling, 3 = High, 4 = Rising, none = no assigned tone), columns = predicted tone.

	0	1	2	3	4	none
0	862	21	65	49	9	3
1	130	309	30	81	6	3
2	281	24	188	54	9	3
3	82	25	21	218	5	1
4	123	19	19	30	133	3
none	0	0	0	0	0	0

Table 7: Thai tone confusion (**Flattened**). Rows = reference tone (0=Mid, 1=Low, 2=Falling, 3=High, 4=Rising, none=no tone), columns = predicted tone.

	1	2	3	4	5	6	none
1	751	32	172	16	11	25	10
2	188	354	42	23	4	46	6
3	73	20	440	68	20	34	7
4	33	58	18	99	21	40	3
5	10	2	29	21	71	7	3
6	35	30	60	26	24	184	4
none	0	0	0	0	0	0	0

Table 8: Vietnamese tone confusion (**Original**). Tones: 1 = mid, 2 = huyền (falling), 3 = sắc (rising), 4 = hỏi (dipping), 5 = ngã (creaky), 6 = nặng (heavy), none = no tone. Rows = reference, columns = predicted.

	1	2	3	4	5	6	none
1	815	7	135	14	13	15	9
2	500	43	78	14	6	13	6
3	332	9	260	13	12	29	5
4	125	3	49	68	14	11	3
5	49	1	14	21	41	16	2
6	91	9	109	22	20	105	5
none	0	0	0	0	0	0	0

Table 9: Vietnamese tone confusion (**Flattened**). Tones: 1=mid, 2=falling, 3=rising, 4=dipping, 5=creaky, 6=heavy, none=no tone. Rows = reference, columns = predicted.

	1	2	3	4	none
1	678	64	42	164	13
2	78	700	97	163	34
3	56	99	434	104	22
4	145	130	97	1192	34
none	12	31	14	26	121

Table 10: Mandarin Chinese tone confusion (**Original**). Tones: 1=high-level, 2=rising, 3=dipping, 4=falling, none=no tone. Rows = reference, columns = predicted.

	1	2	3	4	none
1	553	163	66	154	21
2	207	529	85	200	40
3	130	126	290	134	29
4	451	252	140	703	52
none	26	25	9	25	116

Table 11: Mandarin Chinese tone confusion (**Flattened**). Tones: 1=high-level, 2=rising, 3=dipping, 4=falling, none=no tone. Rows = reference, columns = predicted.

A discovery procedure for synlexification patterns in the world's languages

Hannah S. Rognan
University of Toronto
Department of Linguistics
hannah.rognan@mail.utoronto.ca

Barend Beekhuizen
University of Toronto, Mississauga
Department of Language Studies
barend.beekhuizen@utoronto.ca

Abstract

Synlexification is the pattern of crosslinguistic lexical semantic variation whereby what is expressed in a single word in one language, is expressed in multiple words in another (e.g., French *monter* vs. English *go+up*). We introduce a computational method for automatically extracting instances of synlexification from a parallel corpus at a large scale (many languages, many domains). The method involves debiasing the seed language by splitting up synlexifications in the seed language where other languages consistently split them. The method was applied to a massively parallel corpus of 198 Bible translations. We validate it on a broad sample of cases, and demonstrate its potential for typological research.

1 Introduction

Languages vary in how they 'package' the same conceptual content in words. Variation in colexification – a word in one language having two or more (partial) translation equivalents in another (e.g., English *blue* translating to Russian *sinij* 'dark blue' and *goluboj* 'light blue'), has been widely studied (François, 2008; Östling, 2016; Kemp et al., 2018). Another kind of variation occurs when a word in one language is, on the same occasion, translated as two or more words in another language. For example, French *monter* translates to English *go* and *up*. Here, the complex concept expressed by a single lexical item in one language is split into two constituent concepts in another language – i.c. English *go* expressing 'motion', and *up* the 'vertically elevated' nature of the goal location. While this kind of variation has been studied for individual cases, its generalization was only recently explicated by Haspelmath (2023), who dubbed the phenomenon 'synlexification', and its inverse 'circumlexification' (e.g., French *monter* synlexifies what English *go+up* circumlexifies).

Parallel corpora have been successfully used to investigate crosslinguistic patterns of colexification (Wälchli, 2014; Liu et al., 2023; Beekhuizen et al., 2024). However, extant computational approaches are by design unable to find cases of synlexification. Furthermore, existing corpus-based studies for individual cases do not allow for general discovery across semantic domains, which would be desirable to better understand the determinants of the typological variation in synlexification patterns. Our procedure aims to overcome these challenges.

In this paper, we first review corpus-based studies of synlexification across several semantic domains, motivating a more systematic approach. We then introduce a two-step model for automatically extracting synlexification patterns from parallel corpora. We validate the extracted patterns through comparison with documentary resources (grammars and dictionaries), and show that our method captures both many known and novel cases of synlexification. Finally, we present an initial exploration of the typological variation. Code and (shareable) data are available through https://github.com/dnrb/synlexification.

2 Background

2.1 Synlexifciation across domains

Motion verbs provide a well-established domain for studying synlexification, as languages vary in how they encode the manner of motion ('walking', 'rolling', 'going') and the path ('up', 'out', 'back'). Central here is Talmy (1991)'s distinction between satellite-framed and verb-framed languages. In the former (e.g., Germanic), manner is expressed through the verb and path through a particle, while verb-framed languages (e.g., Romance) encode the path directly in the verb, such as French *monter*, corresponding to *go* and *up* in English. Verkerk (2013) used a parallel corpus of Indo-European languages to examine crosslinguistic variation in

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 93–113
August 1, 2025 ©2025 Association for Computational Linguistics

motion event expression.

Causatives are another domain in which typological differences in synlexification are prevalent (Levshina, 2015). Languages vary in whether lexicalize caused events as single verbs (e.g., *show* 'cause to see') or express them analytically (e.g., with one element expressing 'cause' and another 'see'). It has been found, using parallel and comparable corpora, that there is variation in the degree to which languages express different types of causation (e.g., 'making' vs 'letting'; Levshina, 2016) and different kinds of events (Haspelmath et al., 2014).

Light verbs form a third domain. Samardžić and Merlo (2010) use parallel corpora and word alignment procedures to investigate how English light verb constructions (e.g., *have a laugh*) align with single verbs in German (e.g., *lachen* 'laugh'). Their results reveal that such English constructions frequently map to one-word expressions in German. Nagy T. et al. (2020) extend this approach by automatically detecting cross-linguistic equivalents of light verb constructions in 4 languages. Both papers demonstrate that parallel corpora and word alignment techniques with automated decision procedures can highlight systematic variation in synlexification patterns across languages.

Negative verbs Different strategies for expressing negation have been found in the world's languages (Miestamo, 2007). One way to express negation is to combine a verb with a separate negative marker (e.g., *not+know*), another is to incorporate the negative meaning in a single word such as Tundra Nenets *yexara-* 'not know' (Nikolaeva, 2014, p. 285), and some words are inherently negative like *lack* and *refuse* (Miestamo, 2007). Some languages have been noted to deploy such synlexifying forms more than others (e.g., Ainu; Kwong, 2017), and some semantic domains are more likely to have synlexifying negative verbs (e.g. existentials; Veselinova, 2013).

Compounding, finally, is the morphological strategy of forming new lexical items from other lexical items. Languages vary in the extent to which they apply this strategy or instead choose to 'label' the concept (Štekauer et al., 2012), thus synlexifying what the compounding language circumlexifies. A notable case studied through parallel corpora are co-compounds, which consist of nouns that frequently occur in similar contexts (Wälchli, 2005, 2007), such as *hand-foot* meaning *limbs*.

These studies demonstrate the prevalence of synlexification across domains and languages, and validate the use of parallel corpus methods for identifying such patterns. However, these approaches focused on specific constructions or lexical domains, with top-down methods for detecting instances of the variation. Our approach proposes a bottom-up, scalable extraction method that identifies synlexification patterns across many languages and domains simultaneously, enabling both replication of known patterns and discovery of novel ones.

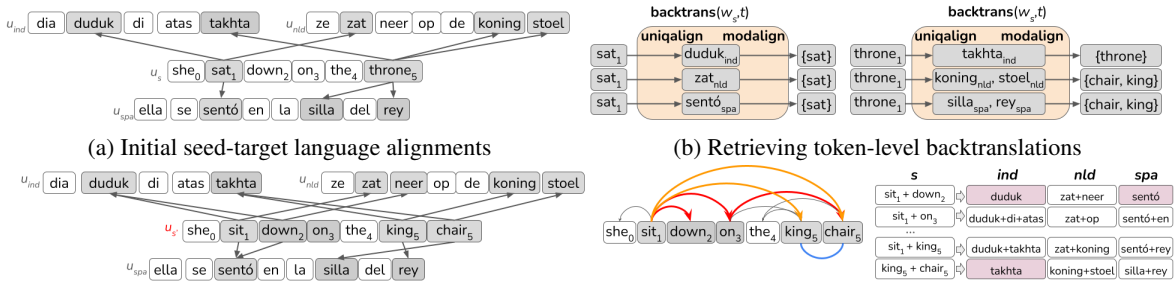
2.2 Explanations of synlexification patterns

Although explanations of the cross-linguistic variation in synlexification patterns has not been studied systematically, Haspelmath (2023) suggests Mańczak (1966)'s law of differentiation as a candidate explanation. This 'law' states that more frequently used meanings are more likely to be differentiated. The intuition is that more frequent groups of concepts are more likely to be synlexified, while less frequent groups of concepts are expected to remain circumlexified. While colexification patterns have been studied along these lines (e.g., Kemp et al., 2018), only initial evidence for the application of this idea to synlexification has been found in the form of the lexical vs. analytic causatives (Haspelmath et al., 2014).

Synlexification patterns are also expected to vary between languages. Ullmann (1966) notes that German tends to use more circumlexified forms than English or French. Aranovich and Wong (2023) distinguish between 'lexicological languages', such as Chinese, which tend to use more lexical items to express complex concepts, and 'grammatical languages', such as Sanskrit, which rely more on grammatical constructions. Seiler (1975) presents a similar distinction, but draws attention to the nature of the semantic operation, with some languages 'describing' (circumlexifying) complex concepts (e.g., Swedish *morbror* 'mother brother' and *farbror* 'father brother') and others 'labelling' (synlexifying) them (e.g. English *uncle*). Our approach can shed light on the extent to which languages as a whole tend to follow certain strategies.

2.3 Goals

To study patterns of synlexification at scale (many languages, many lexical fields), an automated extraction procedure is necessary. Existing automated procedures, all focussing on colexification patterns, include Wälchli (2014); Liu et al. (2023); Viechnicki et al. (2024) and Beekhuizen et al. (2024).



(a) Initial seed-target language alignments

(b) Retrieving token-level backtranslations

(c) Alignments with the granularized seed language

(d) Identifying synlexifications (purple)

Figure 1: Schematic illustration of the synlexification detection model

However, presumably as a means to restrict the search space, none of these procedures consider the alignment of one element in one language with multiple elements in another language, which is the goal of the current study. As such, this paper presents a novel approach that allows for the detection of instances of synlexification in a massively parallel corpus, i.e., a corpus where one text is translated into many languages.

By operationalizing a typological insight about the variable expression of 'the same' meanings through formal means, this paper further aims to contribute to the emerging field of (corpus-based) algorithmic typology (Wälchli, 2014; Wälchli and Sjöberg, 2024), in which typological concepts are investigated through formalization and quantification. This paper explores the notion of synlexification as proposed by Haspelmath (2023) by looking at the linguistic patterns that emerge when it is fairly directly applied to translated text.

After validating the output of this method, we present initial explorations of the kinds of insights this method can lead to: the nature of the variation of the occurrence of synlexification across lexical domains and languages, the discovery of novel domains of synlexification, and the functional determinants of the likelihood of synlexification.

3 A synlexification detection model

To find instances of variation in the lexification of the same concepts, we first have to identify such concepts. Given that translations of the same message can be expected to express more or less the same lexical-semantic content, we can use a bitext B_t between a seed language s and a target language t to provide us with **comparison concepts** (i.e., analytic concepts that allow us to compare languages without making a claim as to their language-specific validity; Haspelmath, 2018), and apply word alignment techniques (Tiedemann, 2011) to determine if there are recurrent many-to-one mappings between multiples of words in s and singleton words in t. This, then allows us to compare across target languages if the same multiples are aligned to singletons across target languages. Step 2 of the method describes this procedure.

However, a seed language may synlexify what target languages circumlexify, and one-to-many alignments from a seed language to a target language are not comparable: if English *enter* aligns with Dutch *ga+binnen* and German *tritt+ein*, we have no way of knowing whether Dutch and German circumlexify the complex concept expressed by *enter* similarly. To maintain the set-up of Step 2 (i.e., finding many-to-one s-to-t mappings), Step 1 creates a synthesized version of the seed language corpus in which seed language words are artificially circumlexified if other languages reliably do so.

3.1 Preliminaries

Several definitions will be used throughout. We define a **bitext** to a target language t as $B_t = [\langle u_s^1, u_t^1\rangle, \langle u_s^2, u_t^2\rangle, \ldots, \langle u_s^n, u_t^n\rangle]$, where u_s^i is a seed language utterance and u_t^i a sentence-aligned target language utterance. **Word types** in a (seed or target) language l are denoted $v_s \in V_s$ while **word tokens** are denoted $w_s \in W_s$. (Types and tokens need to be kept separate for several definitions). The function $\mathbf{type}(w^i)$ retrieves the type v associated with a word token w^i.

Given a set of word alignments between tokens of s and t, derived through some alignment procedure, the function $\mathbf{align}(w_l, l')$, then, retrieves the word alignments between a token $w_l \in u_l$ and a set of tokens $\{w_{l'}^i, w_{l'}^j, \ldots, w_{l'}^n\} \subseteq u_l'$ for a language pair $\langle l, l'\rangle$. A further function, $\mathbf{uniqalign}(w_l, l', Q)$ retrieves tokens aligned to w_l that themselves align exclusively to words in some set Q consisting of word tokens of l.

$\lvert\{w_s \lvert \textbf{type}(w_s) = v_s \wedge P \subseteq \textbf{backtrans}(w_s, t)\}\rvert$	$\lvert\{w_s \lvert \textbf{type}(w_s) = v_s \wedge P \not\subseteq \textbf{backtrans}(w_s, t)\}\rvert$
$\lvert\{w_s \lvert \textbf{type}(w_s) \neq v_s \wedge P \subseteq \textbf{backtrans}(w_s, t)\}\rvert$	$\lvert\{w_s \lvert \textbf{type}(w_s) \neq v_s \wedge P \not\subseteq \textbf{backtrans}(w_s, t)\}\rvert$

Table 1: Four quantities going into the Fisher Exact test to determine $\textbf{association}(v_s, P, t)$.

Formally, $\textbf{uniqalign}(w_l, l', Q) = \{w_{l'} | w_{l'} \in \textbf{align}(w, l') \wedge \textbf{align}(w_{l'}, l)/Q = \emptyset\}$. Two further functions build on the alignment functions. First, $\textbf{modalign}(v_l, l')$ ('modal alignments') returns the word type in l' that is most commonly (modally) aligned with v_l. Formally, $\textbf{modalign}(v_l, l') =$

$$\arg\max_{v_{l'} \in V_{l'}} |\{w_l | \textbf{type}(w_l) = v_l \wedge \exists w_{l'}.\textbf{type}(w_{l'}) = v_{l'} \wedge w_{l'} \in \textbf{uniqalign}(w_l, l')\}|.$$

Second, $\textbf{backtrans}(w_l, l')$ returns the set of modal alignments of the word tokens $\{w^i_{l'}, \ldots, w^n_{l'}\}$ given by $\textbf{uniqalign}(w_l, l', \{w_l\})$, or: the most-common backtranslations into l of w_l, given its alignments to tokens in l' that only align to w_l themselves. To exemplify, given a set of s-t alignments (Figure 1a), the $\textbf{backtrans}$ function (Figure 1b) retrieves the most-commonly aligned word of the target word tokens aligned with each of the seed language tokens.

3.2 Synthesizing a circumlexified seed corpus

We propose that cross-linguistically recurrent, and statistically reliable one-to-many alignments between the seed language s and the various target languages $t \in T$ allow us to replace synlexified concepts in s by synthesized circumlexifications. This procedure requires us to define what alignments are reliable and how to determine which tokens are replaced by a circumlexification.

First, for every seed language word type $v_s \in V_s$ with lexical meaning (here: nouns, adjectives and verbs) we retrieve all significantly associated potential circumlexifications, or: **paraphrases**, where a paraphrase $P = \{v^i_s, \ldots, v^n_s\}$, that is: a set of seed language word types (possibly including v_s itself), requiring $|P| \geq 2$ so that the paraphrase is into *more* words than the original. A paraphrase P is significantly associated with v_s given t, or: $\textbf{association}(P, v_s, t) = \top$, if a Fisher-Exact test over the 2×2 table in Table 1 yields a p-value below a pre-set threshold $\theta_{fe} \in (0, 1)$, and $\textbf{association}(P, v_s, t) = \bot$ otherwise.

Concretely, the Fisher Exact test assesses whether the association between v_s and P given a target language t is significant if the number of tokens of v_s whose modal backtranslations into s include P (top-left cell) is higher than expected by chance, that is: compared to (1) the set of tokens of v_s whose backtranslations do *not* include P (bottom-left cell), and (2) the set of tokens of other types whose backtranslations *do* include P (top-right cell). Following the example of Figure 1, if for Spanish the backtranslation *{king,chair}* occurs across many tokens of English *throne*, and *{king,chair}* infrequently occurs as the backtranslation of other English word types, the association between v_s = *throne* and P = *{king,chair}* is likely significant. Note that we use the inclusion of P in the backtranslation of w_s rather than the identity of P and $\textbf{backtrans}(w_s, t)$, because spurious backtranslations may occur in noisy alignments, thus weakening the $\langle v_s, P\rangle$ associations.

Retrieving significant $\langle v_s, P\rangle$ associations across target languages allows us, then, to leverage the crosslinguistic frequency of such association. If many target languages circumlexify v_s *in the same way* (i.e., backtranslating to the same paraphrase P), we have evidence that relevant tokens of v_s should be replaced by P, so that we would be able to identify that those languages circumlexify what other languages (including the seed language) synlexify. We approach this issue by iteratively replacing the seed language tokens whose types show significant v_s, P associations across the greatest number of target languages, as follows.

First, let $T_{\langle v_s, P\rangle}$ be the set of target languages for which $\textbf{association}(v_s, P, t) = \top$. We define the best word type and paraphrase pair $\langle v^{\max}_s, P^{\max}\rangle = \arg\max_{\langle v_s, P\rangle} |T_{v_s,P}|$, that is: the pair with the greatest number of languages for which it is significant. (Ties between $\langle v_s, P\rangle$ pairs are broken by the average p-value of the Fisher-Exact tests given the languages in $T_{v_s,P}$, prioritizing lower p-values).

Next, given $\langle v^{\max}_s, P^{\max}\rangle$, the set of **replaced tokens** is defined as $\{w_s | \textbf{type}(w_s) = v^{\max}_s \wedge \exists t.(t \in T_{\langle v^{\max}_s, P^{\max}\rangle} \wedge P \subseteq \textbf{backtranslate}(w_s, t))\}$, or: the set of tokens of v_s that backtranslate for at least one target language $t \in T_{\langle v^{\max}_s, P^{\max}\rangle}$ to a set of seed language types that include $P^{\max}$. These tokens are then replaced by $P^{\max}$ in a new corpus of the granularized seed language s', and removed from

W_s, after which the **association** mappings are recomputed. The procedure then repeats, calculating a novel $\langle v_s^{\max}, P^{\max}\rangle$, until $|T_{\langle v_s^{\max}, P^{\max}\rangle}| < \theta_{bt}$, i.e., the set of languages for which the $\langle v_l^{\max}, P^{\max}\rangle$ association is significant is smaller than some pre-set threshold $\theta_{bt} \in [1..\infty]$.

3.3 Finding reliable synlexification patterns

Next, the circumlexified seed language is word-aligned with each of the target languages (Figure 1c), and Step 2 involves finding reliable alignments between word pairs in s and words in t. To constrain the search space, we only consider pairs of seed language word tokens that meet two requirements. First, the pair contains one member with a lexical part of speech (here: nouns, adjectives, verbs, adpositions) and one member with either such a part of speech or a contentful satellite element (derivational affix, adverb, particle, proper noun). Second, the pair consists of elements of the same paraphrase P (blue line in Figure 1d for *king* and *chair*), or stand in a head-dependent relation to each other in a dependency parse (red lines; e.g., *sit* and *down*), including a second-order relation linking heads to the nominal dependents of any adpositions headed by the heads (orange lines; e.g., *sit* and *chair*). The first criterion restricts the search space to only parts of speech expressing lexical content, which is what we are centrally interested in. Second, all attested cases of synlexification (cf. §2) involve elements in a grammatical head-dependency relation to each other in circumlexifying languages, suggesting that this is a reasonable restriction of the search space.

For each pair of word tokens in the granularized seed language $\langle w_{s'}^i, w_{s'}^j\rangle$ meeting these criteria, we now retrieve the alignments in t using $\mathbf{uniqalign}(w_{s'}^i, l', \{w_{s'}^i, w_{s'}^j\})$. The **lexification** function, defined formally as

$$\mathbf{lexification}(w_{s'}^i, w_{s'}^j) = \begin{cases} \text{synlexified} & \text{if } \mathbf{ua}(w_{s'}^i, t, S) \cap \mathbf{ua}(w_{s'}^j, t, S) \neq \emptyset, \\ \text{unlexified} & \text{if } \mathbf{ua}(w_{s'}^i, t, S) = \emptyset \wedge \mathbf{ua}(w_{s'}^j, t, S) = \emptyset, \\ i\text{-lexified} & \text{if } \mathbf{ua}(w_{s'}^i, t, S) = \emptyset \wedge \mathbf{ua}(w_{s'}^j, t, S) \neq \emptyset, \\ j\text{-lexified} & \text{if } \mathbf{ua}(w_{s'}^i, t, S) \neq \emptyset \wedge \mathbf{ua}(w_{s'}^j, t, S) = \emptyset, \\ \text{circumlex.} & \text{otherwise,} \end{cases}$$

(where $\mathbf{ua} = \mathbf{uniqalign}$, $S = \{w_{s'}^i, w_{s'}^j\}$ and 'circumlex.' is short for 'circumlexified') determines the lexification category. A pair of tokens is said to be **synlexified** if both tokens are **uniqalign**-ed to the same token(s) in t, **unlexified** if both tokens **uniqalign** to no words in t, i-**lexified** if $w_{s'}^j$ has no alignments but $w_{s'}^i$ does, j-**lexified** if, conversely, $w_{s'}^i$ has no alignments but $w_{s'}^j$ does, and **circumlexified** otherwise (i.e., both tokens have alignments in u_t but these sets do not overlap).

4 Experimental set-up

Corpus We test our model on a corpus of Bible translations gathered through the `bible.is` API. While this corpus has issues of ecological validity owing to the nature of the concepts expressed (being exogenous to many cultures) and the frequent production of these texts by non-native speakers (Pinhanez et al., 2023; Domingues et al., 2024), it has been used extensively in successfully identifying patterns of crosslinguistic lexical semantic variation that align with observations based on other data sources (Wälchli, 2014; Asgari and Schütze, 2017; Liu et al., 2023). Recognizing the non-identity between the translated, religion-oriented variety of a language and other, more ecologically valid, genres, we use the term **doculect** (Cysouw and Good, 2013) to refer to the variety of a language documented through translation. With these caveats, we treat the results as a lower-bound estimate of the real variation.

Preprocessing A sample of 198 doculects was derived through diversity sampling (Miestamo et al., 2016) ensuring areal and genetic diversity (see Appendix A for a list of doculects). The seed doculect, not part of the sample, was set to be the World English Bible translation. The text of this doculect was preprocessed by lemmatizing, PoS-tagging, and dependency-parsing it with SpaCy (Honnibal and Montani, 2017) and subsequently splitting derivationally complex words (e.g. *un-believe-able*) through CELEX2 (Baayen et al., 1996), as these reflect complex meanings that may be synlexified in other doculects. Target doculects were preprocessed by removing punctuation and segmented with VORM, a state-of-the-art unsupervised canonical morphological segmentation model (Beekhuizen, 2025b), which segments words into stems and affixes. Alignments for both s and s' were subsequently derived with Eflomal (Östling and Tiedemann, 2016), using the 'grow-diag-final-and' heuristic. The alignment in Step 1 was done with stems in the target doculects only (as inclusion of affixes at this state led to noise in the procedure), whereas the alignment for Step 2 also included affixes.

Parameters The significance threshold θ_{fe} was

set at 1^{e-6} and the minimum number of languages for which a the $\langle v_s, P\rangle$ association was significant was set as $\theta_{bt} = 3$. Both values were based on post-hoc assessment of the extraction quality and more complete parameter tuning on a benchmark set will be left for future research. Among the valid circumlexified seed language word pairs, only those that occurred ≥ 10 times throughout the circumlexified seed data were kept for further analysis.

5 Validating the model

Step 1 of the method splits 896 vocabulary types in the seed language, including some cases that are very frequently split among the target doculects, such as *answer=say+answer* (100 doculects) and *smoke=smoke+fire* (60 doculects). Notably, most splits involve splitting a word type into the word type itself and an additional element, though cases like *sail=go+boat* (34 doculects) are found as well. A larger sample is presented in Appendix B with the full data being available in the repository.

Based on the circumlexified seed doculect corpus, a total of $2,563$ comparison meaning pairs with a frequency ≥ 10 were found, and alignment patterns into the 198 target doculects were extracted with Step 2 of the extraction algorithm. While the next section demonstrates what can be done with these data, here we first provide a post-hoc validation of the model.

5.1 Validating extracted synlexifications

Given that no evaluation set is available, we validate the model by inspecting its extractions for several well-known cases, alongside several hand-picked ones representing frequently and infrequently synlexified meanings.[1] For each case, we selected one doculect whose predicted most-common strategy was to synlexify the meanings, one that most commonly circumlexified them, and one that most commonly left one meaning underspecified (i or j-lexified). For each doculect, we compared the extracted markers against grammars and dictionaries, referring to the translation tokens to validate. The fields and a qualitative description of the assessment can be found in Appendix C, while Table 2 summarizes the results. Although we had equal numbers of each predicted strategy initially, we conducted the evaluation multiple times as we implemented improvements to the model, which lead to a new strategy prediction for some cases. In addition, the model labeled $3/47$ inspected pairs as dominantly 'unlexified', which means no prediction regarding the modal type could be made. Overall, 80% of the 41 cases that were determined with certainty were correctly labeled by the model.

predicted strategy	correct	uncert.	incorrect
Synlexified (13)	0.85	0.15	0.00
Circumlexified (15)	0.80	0.00	0.20
x-lexified (16)	0.63	0.06	0.31

Table 2: Results of manual validation. 'uncert.'=uncertain; (N) = number of cases.

Among the accurate cases, we find the pair *enter+in*, synlexified as *natt* in Fulfulde (cf. McIntosh, 1984, p. 125: *natt-ay* 'enter'), circumlexified as *go+iin* in Jamaican English (cf. Bailey, 1968, p. 227), and j-lexified in Karkar as *mek* (cf. Rigden, n.d., p. 112, 116: *mek* 'in'). The majority of errors were i or j-lexifications which should have been labeled as cases of circumlexification. For instance, the pair *to+world* was labeled as underspecified in Bora because the pair frequently aligns only to *ííñujɨ́* (*land*) and in other cases to *-vu* (a spatial goal marker; Thiesen and Weber, 2012, p. 156), but the pair should have been aligned to both of these words to indicate circumlexification – an error attributable to the strictness of the **uniqalign** procedure, as many instances of these markers were found to have spurious alignments. Another type of error involved doculects synlexifying a concept but predicted to circumlexify. For instance, Ndyuka synlexifies the pair *un-+clean* with the word *takuu* meaning 'evil' (Huttar and Huttar, 1994, p. 62), but the model defines the tokens as circumlexified because in many instances, *takuu* is aligned with only one member of the pair ⟨*un-*, *clean*⟩, and other Ndyuka words with the other member.

6 Exploring synlexification patterns

The validation suggests that the method is a reasonable first attempt at extracting patterns of synlexification at a lexicon-wide scale and for a typologically diverse sample of doculects. Next, we explore applications of the extracted data, to demonstrate the linguistic use of the method.

[1] Notably, this validation step provides more evidence for the quality of the extraction than most other computational methods (e.g. Liu et al., 2023; Beekhuizen et al., 2024), though see Beekhuizen (2025a) and Beekhuizen (2025c) for thoroughly evaluated extraction algorithms.

part of speech pair	N	% syn.	top-3 most frequently synlexified (N doculects)
Adposition+Noun	461	18%	mountain+on (116) before+foot (80) in+peace (72)
Adposition+Verb	460	19%	rise+up (107) get+up (104) down+fall (95)
Noun+Verb	408	38%	bread+eat (173) law+write (164) apostle+send (163)
Verb+Verb	245	20%	deceive+lie (162) suffer+torment (146) persecute+suffer (120)
Noun+Noun	198	81%	boat+ship (186) boat+sea (186) horse+soldier (184)
Adjective+Noun	87	68%	blind+eye (179) blood+dead (178) famine+hungry (172)
Affix+Verb	86	90%	teach+-er (142) serve+-ant (105) pray+-er (105)
Proper Noun+Verb	84%	10%	Peter+answer (2) Jesus+answer (1) Christ+die (1)

Table 3: Synlexification across PoS pairs. N= number of pairs, % syn.= % of pairs synlexified in ≥ 1 doculect.

Distribution across the lexicon. Most (1715/2563, or 67%) comparison concept pairs are not dominantly lexified in any doculect (where 'dominant' means 'applied in $\geq 50\%$ of the tokens of that pair'). This suggests that synlexification happens in select areas of the lexicon. Breaking down the pairs by their grammatical categories (Table 3; a larger sample is given in Table 10 in App. D), we find substantial variation: combinations of proper nouns and verbs are for instance rarely synlexified (9%). Conversely, many of the noun+noun, adjective+noun, and affix+noun pair have at least one doculect synlexifying them, possibly due to such combinations building complex categories that can variably be 'described' or 'labelled' (cf. Seiler, 1975) for communication to succeed.

Motion verb synlexification, part of the preposition+verb combinations, can be found among the most frequently dominantly synlexified preposition+verb pairs – e.g., *rise+up* or *down+fall*, but other preposition+verb combinations reflect more 'accidental' combinations of verbs and prepositions, making preposition+verb pairs have a low number of synlexifying doculects. While looking at the level of grammatical categories is likely too coarse a subdivision, the variation across grammatical categories suggests that some of the uneven distribution across the lexicon may be related to the types of concepts they denote.

Areal distribution. Secondly, not all doculects are equally likely to synlexify, as discussed in §2. There are substantial areal patterns, with the average number of comparison meaning pairs dominantly synlexified ranging from 122 (Australian doculects), and 150 (South-America), over 173 (North-America) and 190 (Papunesia), to 215 (Africa) and 228 (Eurasian). Figure 2 plots the number of dominantly synlexified pairs across the 198 doculects. These areal patterns are open to multiple interpretations. The high numbers for the European doculects (Basque, Dutch, Finnish, Hungarian, Greek) might reflect the extended exposure of these cultures to the cultural concepts of Christianity ('pray', 'temple', 'prophet', ...), leading to short, synlexifying forms. However, not all variation can be attributed to cultural factors, as there is substantial variation between other macro-areas where the dissemination of these religious concepts is more recent. Moreover, the clearly religious concepts form only a small subset of all variably synlexified concepts.

Potential for case studies. Synlexification patterns have mostly been studied for specific semantic domains (cf. §2). The proposed procedure allows us to study such cases by retrieving matching comparison concept pairs. The well-studied case of **motion events** can for instance be studied by looking for motion verbs (*go, fall, sit, put, ...*) and particles (*in, out, up, down, ...*). For most such pairs, which are presented in Table 11 in Appendix E, doculects do not synlexify. Most frequently synlexified are five pairs of motion along the vertical axis: *rise+up* ($N = 107$), *get+up* ($N = 104$), *fall + down* ($N = 95$), *sit + down* ($N = 81$), and *stand + up* ($N = 67$). Notably, in some of these cases the direction of movement is already implicated by the manner of motion verb. These cases raises interesting questions about the concept of synlexification per se. If one language l circumlexifies this complex concept into a pair of lexical items $v_l^{\text{fall}}, v_l^{\text{down}}$, aligning with 'fall' and 'down', and another language l' synlexifies them with one lexical item $v_{l'}$, which dictionaries define as 'fall', does this mean that the meanings of v_l^{fall} and $v_{l'}$ are different with the former underspecifying the 'down' component? This seems counterintuitive: after all, even in a synlexifying language like English, *He fell* (as opposed to *He fell down*) at least implicates and perhaps entails 'down'.

Conversely, several of the cases for which sub-

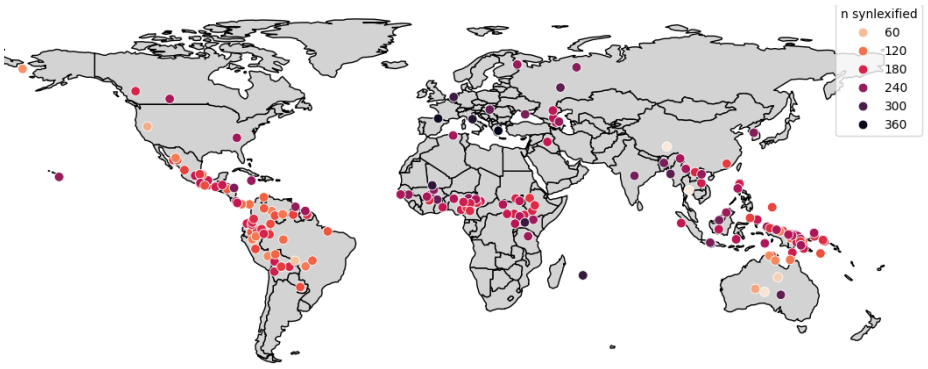


Figure 2: Areal distribution of the number of comparison meanings that are dominantly synlexified, per doculect

stantial typological variation is expected (*enter* + *in* and *go* + *out*) were dominantly synlexified only rarely across doculects ($N = 3$ resp. $N = 2$). While further validation and linguistic analysis is necessary, these data suggest that matters are more complex than the simple path vs. satellite-framing typology lets on.

The bottom-up discovery procedure further allows us to explore new domains involving variation in the synlexification patterns. Speech events form one such domain: **manner of speech verbs**, such as *promise*, *lie*, *answer* and *ask* are often found synlexified, like in English, but more frequently (across doculects) circumlexified into an element translating to English *say* and another to the manner (i.c., *promise*, *false*, *answer*, *ask*). Table 12 in Appendix E presents an overview. While typological observations about speech verbs have been made for small sets of languages (Caballero and Paradis, 2017), the method presented here supports a larger-scale typological comparison.

Mańczak's law of differentiation. Finally, we explore the hypothesis that more frequent meaning pairs are more likely to be synlexified, due to communicative efficiency. We evaluate this hypothesis with the following logistic regression model:

$$\begin{aligned}\text{synlexified} \sim\ & \log \text{pair.frequency} + \text{pos} + \\ & \text{macroarea} + (1|\text{doculect}) + (1|\text{pair})\end{aligned}$$

That is: for each doculect and for each pair, we predict whether the doculect dominantly synlexifies the pair on the basis of the log-frequency of the comparison meaning pair, as derived in Step 2, the part of speech ('pos'; dummy-coded for the 5 most frequent parts of speech pairs, with other pos-pairs coded as 'other'), and the macroarea (dummy-coded). Random intercepts were added for doculects and pairs, reflecting biases of individual doculects or pairs that should be included to constrain the inferred effects of the target variables.

Table 13 in App. F presents full regression results. Critically, over and above significant effects of 'pos' and 'macroarea', the frequency of the meaning pair significantly predicts the likelihood of that pair being synlexified, with the positive direction being in line with Mańczak's law of differentiation. The effect size is furthermore substantial: the observed log Odds Ratio of 1.203 means that for every unit increase in log frequency (e.g, going from $\log N = 3$ to $\log N = 4$, or: $N \approx 20$ to $N \approx 54$), the the likelihood of synlexifying the pair increases more than threefold ($\exp 1.203 \approx 3.330$). Two concerns here are whether the variably-lexified comparison concepts have enough ecological validity and whether the counts of the English-based comparison meaning pairs are a valid measure of meaning frequency. Addressing these would be paramount to further research.

7 Conclusion

This paper introduced a novel method for extracting patterns of synlexification from a parallel corpus at the scale of 198 languages and the full lexicon and validated it on over 40 cases. While the model performed generally well, substantial room for improvement remains. First, replacing seed language words by other seed language words in Step 1 means that the (co)lexification pattern of the seed language still affects what alignments are likely to be made. Explorations of methods that infer latent discrete n-tuples (e.g., through topic modelling, cf. Blei and Lafferty, 2009) prove difficult to tune to yield desired results. In future work, we hope to develop such improvements, create more rigorous methods of evaluation, and apply the method to more ecologically valid corpora.

References

Faruk Abu-Chacra. 2007. *Arabic: An Essential Grammar*. Routledge, Milton Park, Abingdon, Oxon; New York, NY.

Raúl Aranovich and Alan Wong. 2023. Saussure's cours and the monosyllabic myth: the perception of chinese in early linguistic theory. *Language & History*, 66(1):59–79.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. *arXiv preprint arXiv:1704.08914*.

Gorka Aulestia. 1989. *Basque-English dictionary*. Basque series. University of Nevada Press, Reno.

Yves Avril. 2006. *Parlons Komi*. Parlons. L'Harmattan, Paris.

Nicholas Awde and Muhammad Galaev. 2014. *Chechen-English, English-Chechen Dictionary and Phrasebook*. Routledge, Abingdon, Oxon and New York, NY. Originally published by Curzon Press Ltd, 1997.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. The celex lexical database (cd-rom).

Beryl Bailey. 1968. *Jamaican Creole Language Course*. Peace Corps, Washington, D.C.

Barend Beekhuizen. 2025a. Spatial relation marking across languages: extraction, evaluation, analysis. In *29th Conference on Computational Natural Language Learning (CoNLL 2025)*.

Barend Beekhuizen. 2025b. VORM: Translations and a constrained hypothesis space support unsupervised morphological segmentation across languages. In *29th Conference on Computational Natural Language Learning (CoNLL 2025)*.

Barend Beekhuizen. 2025c. Token-level semantic typology without a massively parallel corpus. In *The 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

Barend Beekhuizen, Maya Blumenthal, Lee Jiang, Anna Pyrtchenkov, and Jana Savevska. 2024. Truth be told: a corpus-based study of the cross-linguistic colexification of representational and (inter)subjective meanings. *Corpus Linguistics and Linguistic Theory*, 20(2):433–459.

Keith Berry and Christine Berry. 1999. *A Description of Abun: A West Papuan Language of Irian Jaya*, volume 115 of *Pacific Linguistics: Series B*. Research School of Pacific and Asian Studies, Australian National University, Canberra.

David M Blei and John D Lafferty. 2009. Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC.

Ross Bowden. 1997. *A Dictionary of Kwoma: A Papuan Language of North-East New Guinea*, volume 134 of *Pacific Linguistics: Series C*. Research School of Pacific and Asian Studies, Australian National University, Canberra.

David Briley. 1997. Four grammatical marking systems in bauzi. In Karl J. Franklin, editor, *Papers in Papuan Linguistics No. 2*, pages 1–131. Research School of Pacific and Asian Studies, Australian National University, Canberra.

Rosario Caballero and Carita Paradis. 2017. Verbs in speech framing expressions. *Journal of Linguistics*, 22(1):1–40.

Eugene H. Casad. 2012. Cora–spanish lexical database (draft). Manuscript. Unfinished lexical database, posted as is without peer review.

Rodolfo Cerrón-Palomino. 2006. *El Chipaya o Lengua de los Hombres del Agua*, 1 edition. Fondo Editorial, Pontificia Universidad Católica del Perú, Lima.

Michael Cysouw and Jeff Good. 2013. Languoid, doculect and glossonym: Formalizing the notion 'language'. *LANGUAGE DOCUMENTATION & CONSERVATION*, 7.

Rudolf Pieter Gerardus de Rijk. 2008. *Standard Basque: A Progressive Grammar*. MIT Press, Cambridge, MA.

René Dirven, Louis Goossens, Yvan Putseys, and Emma Vorlat. 1982. *The scene of linguistic action and its perspectivization by speak, talk, say and tell*. John Benjamins.

Pedro Henrique Domingues, Claudio Santos Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. Quantifying the ethical dilemma of using culturally toxic training data in AI tools for indigenous languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 283–293, Torino, Italia. ELRA and ICCL.

Norm Draper and Sheila Draper. 2002. *Dictionary of Kyaka Enga: Papua New Guinea*, volume 532 of *Pacific Linguistics*. Pacific Linguistics, Canberra.

Tom E. Dutton and Dicks Thomas. 1994. *A new course in Tok Pisin (New Guinea Pidgin)*, volume 67 of *Pacific Linguistics : Series D, Special Publications*. Australian National University, Canberra.

Irwin Firchow. 1974. Rotokas grammar. Manuscript. Accessed via SIL International.

Alexandre François. 2008. Semantic maps an the typology of colexifications: Intertwining polysemous networks across languages. In Martine Vanhove, editor, *From polysemy to semantic change: Towards a typology of lexical semantic associations*, pages 163–216. John Benjamins, Amsterdam.

Martin Haspelmath. 2018. How comparative concepts and descriptive linguistic categories are different. In Daniël Olmen, Tanja Mortelmans, and Frank Brisard, editors, *Aspects of linguistic variation*, pages 83–114. De Gruyter Mouton.

Martin Haspelmath. 2023. Coexpression and synexpression patterns across languages: comparative concepts and possible explanations. *Frontiers in Psychology*, 14.

Martin Haspelmath, Andreea Calude, Michael Spagnol, Heiko Narrog, and Elif Bamyaci. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of linguistics*, 50(3):587–625.

Steffen Haurholm-Larsen. 2016. *A Grammar of Garifuna*. Ph.D. thesis, Universität Zürich, Bern.

Jeffrey Heath. 1999. *A Grammar of Koyra Chiini: The Songhay of Timbuktu*. Number 19 in Mouton Grammar Library. Mouton de Gruyter, Berlin and New York.

Jeffrey Heath. 2014. A grammar of yorno-so. Draft manuscript, November 2014.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

George L. Huttar and Mary L. Huttar. 1994. *Ndyuka*. Descriptive Grammars Series. Routledge, London and New York.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1):109–128.

Yan Kit Kwong. 2017. Lexical negative verbs in ainu language. *International Journal of Humanities and Social Science*, 7(2):166–170.

Natalia Levshina. 2015. European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica*, 49(2):487–520.

Natalia Levshina. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen european languages. *Folia Linguistica*, 50(2):507–542.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schütze. 2023. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12969–13000, Toronto, Canada. Association for Computational Linguistics.

Arnold Lock. 2011. *Abau Grammar*, volume 57 of *Data Papers on Papua New Guinea Languages*. SIL-PNG Academic Publications, Papua New Guinea. Summer Institute of Linguistics.

Martin Maiden and Cecilia. Robustelli. 2013. *A reference grammar of modern Italian*, 2nd ed. edition. HRG. Routledge, London ;. Includes bibliographical references and index.

Witold Mańczak. 1966. La nature du supplétivisme. *Linguistics*, 4(28):82–89.

Mary McIntosh. 1984. *Fulfulde Syntax and Verbal Morphology*. KPI, in association with University of Port Harcourt Press, London and Boston.

Matti Miestamo. 2007. Negation – an overview of typological research. *Language and Linguistics Compass*, 1(5):552–570.

Matti Miestamo, Dik Bakker, and Antti Arppe. 2016. Sampling for variety. *Linguistic Typology*, 20(2):233–296.

István Nagy T., Anita Rácz, and Veronika Vincze. 2020. Detecting light verb constructions across languages. *Natural Language Engineering*, 26(3):319–348.

Irina Nikolaeva. 2014. *A Grammar of Tundra Nenets*.

Colleen Alena O'Brien. 2018. *A Grammatical Description of Kamsá: A Language Isolate of Colombia*. Ph.D. thesis, University of Hawai'i at Mānoa, Honolulu.

Asmah Haji Omar. 1969. *The Iban language of Sarawak: a grammatical description*. Ph.D. thesis, University of London.

Robert Östling. 2016. Studying colexification through massively parallell corpora. In Paeivi Juvonen Maria Koptjevskaja-Tamm, editor, *The lexical typology of semantic shifts*, chapter 6, pages 157–176. De Gruyter Mouton.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Helma Pasch. 2007. Grammar of location and motion in zande. *APAL (Annual Publication in African Linguistics)*, 5.

Claudio S. Pinhanez, Paulo Cavalin, Marisa Vasconcelos, and Julio Nogima. 2023. Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6174–6182. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

David J. Prentice. 1971. *The Murut Languages of Sabah*. Number 18 in Pacific Linguistics: Series C. Research School of Pacific and Asian Studies, Australian National University, Canberra.

Veda Rigden. n.d. Karkar grammar essentials. Unpublished manuscript, SIL.

Danilo Salamanca. 1988. *Elementos de gramática del Miskito*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA. Ph.D. dissertation, 380pp.

Tanja Samardžić and Paola Merlo. 2010. Cross-lingual variation of light verb constructions: Using parallel corpora and automatic alignment for linguistic research. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 52–60, Uppsala, Sweden. Association for Computational Linguistics.

Arden G. Sanders and Joy Sanders. 1994. Kamasau (wand tuan) grammar: Morpheme to discourse. Unpublished manuscript. Available at: `http://www.sil.org/pacific/png/abstract.asp?id=47683`.

Hansjakob Seiler. 1975. Die prinzipien der deskriptiven und etikettierenden benennung. In Hansjakob Seiler, editor, *I. I. I. Linguistic Workshop*, pages 2–57. Fink, München.

H. L. Shorto. 2013. *Wa-Praok Vocabulary*, volume 6 of *Asia-Pacific Linguistics*. Asia-Pacific Linguistics, Canberra.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C. Ewing. 2010. *Indonesian Reference Grammar*, 2 edition. London: Allen Unwin, St Leonards.

Alan M. Stevens and A. Ed Schmidgall-Tellings. 2010. *A Comprehensive Indonesian–English Dictionary*, 2nd edition. Ohio University Press, Athens, Ohio.

Leonard Talmy. 1991. *Path to realization: A typology of event conflation*. Berkeley Linguistics Society, Berkeley, California. 2010.

Wesley Thiesen and David Weber. 2012. *A Grammar of Bora with Special Attention to Tone*. Number 148 in SIL International Publications in Linguistics. SIL International, Dallas, Texas.

Jörg Tiedemann. 2011. *Bitext alignment*, volume 4. Morgan & Claypool Publishers.

Stephen Ullmann. 1966. Semantic universals. In Joseph H. Greenberg, editor, *Universals of Language*, 2nd edition, pages 217–262. MIT Press, Cambridge, MA.

Annemarie Verkerk. 2013. Scramble, scurry and dash: The correlation between motion event encoding and manner verb lexicon size in indo-european. *Language Dynamics and Change*, 3(2):169 – 217.

Ljuba Veselinova. 2013. Negative existentials: A cross-linguistic study. *Rivista di Linguistica*, 25(1):107–145.

Peter Viechnicki, Kevin Duh, Anthony Kostacos, and Barbara Landau. 2024. Large-scale bitext corpora provide new evidence for cognitive representations of spatial terms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1089–1099, St. Julian's, Malta. Association for Computational Linguistics.

Bernhard Wälchli. 2014. Algorithmic typology and going from known to similar unknown categories within and across languages. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, pages 355–393. Walter de Gruyter.

Earle Waugh, Arok Wolvengrey, Loretta Pete, Ramona Washburn, and Intellinet Team. 2025. Online cree dictionary. Developed by Miyo Wahkohtowin Community Education Authority. Includes lesson plan builder and syllabic content contributions by Ramona Washburn. Accessed: 2025-04-18.

Wedhawati, Marsono, Edi Setiyanto, Dirgo Sabariyanto, Syamsul Arifin, Sumadi, Restu Sukesti, Herawati, Sri Nardiati, Laginem, and Wiwin Erni Siti Nurlina. 2001. *Tata Bahasa Jawa Mutakhir*. Pusat Bahasa Departemen Pendidikan Nasional, Jakarta.

Bernhard Wälchli. 2005. *Co-compounds and Natural Coordination*. Oxford Studies in Typology and Linguistic Theory. Oxford University Press, Oxford.

Bernhard Wälchli. 2007. *Lexical classes: A functional approach to "word formation"*, pages 153–178. De Gruyter Mouton, Berlin, New York.

Bernhard Wälchli and Anna Sjöberg. 2024. A law of meaning. *Linguistic Typology at the Crossroads*, 4(2):1–71.

Pavol Štekauer, Salvador Valera, and Lívia Kőrtvélyessy. 2012. *Word-formation in the world's languages: A typological survey*. Cambridge University Press.

A Data overview

Tables 4-7 present the 198 doculects, along with their affiliation and macro-area.

B Results from Step 1

Table 8 presents a selection form the output of Step 1 of the model as applied to the sample of Bible data. The first 20 extractions ($\langle v_s, P \rangle$ pairs) and every 30th extraction are printed, along with the number of doculects for which this pair was found to be significantly associated ($\mathbf{association}(v_s, P, t)$), their average p-value (negative-$\log_n$ transformed), the number of tokens in the corpus this applies to, and the proportion of all tokens of v_s this number makes up.

C Detailed validation

This section includes details of all the extracted cases of synlexification, circumlexification, and underspecification that we inspected manually using dictionaries and grammars. Table 9 shows each of the pairs of seed words that we looked at, and the strategy that the model predicts for each language, along with the most frequently extracted tokens for the pair and the glosses.

D Fuller table with extracted cases

Table 10 presents a larger set of examples of synlexifications of the different pairs of grammatical categories.

E Typological frequencies for two semantic domains

This section reports on the frequency with which doculects dominantly synlexify sets of comparison meaning pairs. Table 11 shows instances of (caused) motion events. These were based on all cases where one of the verbs *get*, *rise*, *sit*, *go*, *come*, *enter*, *put*, *throw*, *stand*, *depart*, *ascent*, *fall*, *cast*, or *pour* was combined with an adposition/particle from among *in, out, on, off, from, to, back, up, down*. Table 12 shows instances of speech events, based on all pairs where one of the four main speech verbs (Dirven et al., 1982) *say, tell, speak, talk* was combined with any other element. Note that the $N - 175$ instances of comparison meaning pairs including one of those verbs but not synlexified in any language were omitted from the table.

F Regression analysis of synlexification

This section provides further experimental detail on the logistic regression reported in Section 6. Table 13 presents the output of a mixed effects logistic regression (using the `glmer` library in `R`).

ISO 639-3	name	family	macroarea
AAUWBT	Abau	Sepik	Papunesia
ACMAS3	Gilit Mesopotamian Arabic	Afro-Asiatic	Eurasia
ACUTBL	Achuar-Shiwiar	Chicham	South America
AGGPNG	Angor	Senagi	Papunesia
AGMWBT	Angaataha	Angan	Papunesia
ALYXXX	Alyawarr	Pama-Nyungan	Australia
AMFSIM	Hamer-Banna	South Omotic	Africa
AMKWBT	Ambai	Austronesian	Papunesia
AMMWBT	Ama (Papua New Guinea)	Left May	Papunesia
AMNPNG	Amanab	Border	Papunesia
AMPWBT	Alamblak	Sepik	Papunesia
AMRTBL	Amarakaeri	Harakmbut	South America
AMUMVR	Guerrero Amuzgo	Otomanguean	North America
AOJFIL	Mufian	Nuclear Torricelli	Papunesia
ARLTBL	Arabela	Zaparoan	South America
AVAANT	Avar	Nakh-Daghestanian	Eurasia
AVTWBT	Au	Nuclear Torricelli	Papunesia
AZZTBL	Highland Puebla Nahuatl	Uto-Aztecan	North America
BBOBSM	Northern Bobo Madaré	Mande	Africa
BDHWBT	Baka (South Sudan)	Central Sudanic	Africa
BFDWBT	Bafut	Atlantic-Congo	Africa
BIBWBT	Bissa	Mande	Africa
BKLLAI	Berik	Tor-Orya	Papunesia
BOATBL	Bora	Boran	South America
BORWYI	Bororo	Bororoan	South America
BRUNXB	Eastern Bru	Austroasiatic	Eurasia
BSCWBT	Bassari-Tanda	Atlantic-Congo	Africa
BVRXXX	Burarra	Maningrida	Australia
BVZYSS	Bauzi	Geelvink Bay	Papunesia
BYRWBT	Baruya	Angan	Papunesia
BYXWBT	Qaqet	Baining	Papunesia
CABNVS	Garifuna	Arawakan	North America
CAPSBB	Chipaya	Uru-Chipaya	South America
CASNTM	Mosetén-Chimané	isolate	South America
CAXSBB	Lomeriano-Ignaciano Chiquitano	Chiquitano	South America
CBITBL	Cha'palaa	Barbacoan	South America
CBTTBL	Shawi	Cahuapanan	South America
CCOTBL	Comaltepec Chinantec	Otomanguean	North America
CHEIBT	Chechen	Nakh-Daghestanian	Eurasia
CHRPDV	Cherokee	Iroquoian	North America
CJPTJV	Cabécar	Chibchan	North America
CMEWBT	Cerma	Atlantic-Congo	Africa
CONWBT	Cofán	isolate	South America
CRHIBT	Crimean Tatar	Turkic	Eurasia
CRKWCV	Plains Cree	Algic	North America
CRNWBT	El Nayar Cora	Uto-Aztecan	North America
CRXWYI	Central Carrier	Athabaskan-Eyak-Tlingit	North America
CSKATB	Jola-Esulalu	Atlantic-Congo	Africa

Table 4: Overview of doculects used, along with their affiliation and macro-area (Table 1/4).

ISO 639-3	name	family	macroarea
CTGBSB	Chittagonian	Indo-European	Eurasia
DESWBT	Desano	Tucanoan	South America
DIDWBT	Didinga	Surmic	Africa
DIFXXX	Dieri	Pama-Nyungan	Australia
DJKWBT	Aukan	Indo-European	South America
DTSABM	Toro So Dogon	Dogon	Africa
DUDWYI	Hun-Saare	Atlantic-Congo	Africa
ELLELL	Modern Greek	Indo-European	Eurasia
ESEE06	Ese Ejja	Pano-Tacanan	South America
ESSWYI	Central Siberian Yupik	Eskimo-Aleut	Eurasia
EUSNLT	Basque	isolate	Eurasia
FRDWBT	Fordata	Austronesian	Papunesia
FUVLTBL	Hausa States Fulfulde	Atlantic-Congo	Africa
GAHPNG	Alekano	Nuclear Trans New Guinea	Papunesia
GBILAI	Galela	North Halmahera	Papunesia
GHSPNG	Guhu-Samane	Nuclear Trans New Guinea	Papunesia
GRTBBS	Garo	Sino-Tibetan	Eurasia
GUCTBL	Wayuu	Arawakan	South America
GUHWBT	Sikuani	Guahiboan	South America
GUKBSE	Northern Gumuz	Gumuz	Africa
GUPXXX	Bininj Kun-Wok	Gunwinyguan	Australia
HADLAI	Hatam	Hatam-Mansim	Papunesia
HAKTHV	Hakka Chinese	Sino-Tibetan	Eurasia
HTOWBT	Minica Huitoto	Huitotoan	South America
HUNK90	Hungarian	Uralic	Eurasia
HUVTBL	San Mateo del Mar Huave	Huavean	North America
HWCWYI	Hawai'i Creole English	Indo-European	Papunesia
IANPNG	Iatmul	Ndu	Papunesia
IBATIV	Iban	Austronesian	Papunesia
IFBTBL	Batad Ifugao	Austronesian	Papunesia
INDASV	Standard Indonesian	Austronesian	Papunesia
IRKBST	Iraqw	Afro-Asiatic	Africa
ITAR27	Italian	Indo-European	Eurasia
IZZTBL	Izi	Atlantic-Congo	Africa
JAMBSW	Jamaican Creole English	Indo-European	North America
JAVNRF	Javanese	Austronesian	Papunesia
JBUIBS	Jukun Takum	Atlantic-Congo	Africa
JICWBT	Tol	Jicaquean	North America
KABCEB	Kabyle	Afro-Asiatic	Africa
KBHWBT	Camsá	isolate	South America
KERABT	Kera	Afro-Asiatic	Africa
KFBNTA	Northwestern Kolami	Dravidian	Eurasia
KGRLAI	Abun	isolate	Papunesia
KHGNTV	Khams Tibetan	Sino-Tibetan	Eurasia
KHQBIV	Koyra Chiini Songhay	Songhay	Africa
KIAWBT	Kim	Atlantic-Congo	Africa
KMOWBT	Kwoma	Sepik	Papunesia
KMSPNG	Kamasau	Nuclear Torricelli	Papunesia
KNJSBI	Akateko	Mayan	North America
KORSYS	Korean	Koreanic	Eurasia

Table 5: Overview of doculects used, along with their affiliation and macro-area (Table 2/4).

ISO 639-3	name	family	macroarea
KPVIBT	Komi-Zyrian	Uralic	Eurasia
KPWPNG	Kobon	Nuclear Trans New Guinea	Papunesia
KRLNEW	Karelian	Uralic	Eurasia
KRSWYI	Kresh-Woro	Kresh-Aja	Africa
KTOWBT	Kuot	isolate	Papunesia
KYCPNG	Kyaka	Nuclear Trans New Guinea	Papunesia
LEFTBL	Lelemi	Atlantic-Congo	Africa
LMEABT	Peve	Afro-Asiatic	Africa
MAKLAI	Makasar	Austronesian	Papunesia
MBCWBT	Macushi	Cariban	South America
MCAWBT	Maca	Matacoan	South America
MDYBSE	Male (Ethiopia)	Ta-Ne-Omotic	Africa
MEJTBL	Meyah	East Bird's Head	Papunesia
MFEBSM	Morisyen	Indo-European	Africa
MFYWBT	Mayo	Uto-Aztecan	North America
MHIBSU	Ma'di	Central Sudanic	Africa
MHRIBT	Eastern Mari	Uralic	Eurasia
MIFWBT	Mofu-Gudur	Afro-Asiatic	Africa
MILTBL	Peñoles Mixtec	Otomanguean	North America
MIQSBN	Mískito	Misumalpan	North America
MLPTBL	Bargam	Nuclear Trans New Guinea	Papunesia
MOPWBT	Mopán Maya	Mayan	North America
MORBSS	Moro	Heibanic	Africa
MPMTBL	Yosondúa Mixtec	Otomanguean	North America
MPTWBT	Mian	Nuclear Trans New Guinea	Papunesia
MSYPNG	Aruamu	Ramu	Papunesia
MTOTBL	Totontepec Mixe	Mixe-Zoque	North America
MWWHDV	Hmong Daw	Hmong-Mien	Eurasia
MZMWBT	Mumuye	Atlantic-Congo	Africa
NABWBT	Southern Nambikuára	Nambiquaran	South America
NAFWBT	Nabak	Nuclear Trans New Guinea	Papunesia
NASPNG	Naasioi	South Bougainville	Papunesia
NHXNFB	Isthmus-Mecayapan Nahuatl	Uto-Aztecan	North America
NIAIBS	Nias	Austronesian	Papunesia
NIJLAI	Ngaju	Austronesian	Papunesia
NLDHSV	Dutch	Indo-European	Eurasia
NOAWBT	Woun Meu	Chocoan	South America
NTJXXX	Ngaanyatjarra	Pama-Nyungan	Australia
NTPTBL	Northern Tepehuan	Uto-Aztecan	North America
NUYXXX	Wubuy	Gunwinyguan	Australia
OPMTBL	Oksapmin	Nuclear Trans New Guinea	Papunesia
OTQTBL	Querétaro Otomi	Otomanguean	North America
PADWBT	Paumari	Arawan	South America
PAMPBS	Pampanga	Austronesian	Papunesia
PAONAB	Northern Paiute	Uto-Aztecan	North America
PAUPAL	Palauan	Austronesian	Papunesia
PBBDYU	Páez	isolate	South America
PJTXXX	Pitjantjatjara	Pama-Nyungan	Australia
POEWBT	San Juan Atzingo Popoloca	Otomanguean	North America
POIWBT	Highland Popoluca	Mixe-Zoque	North America

Table 6: Overview of doculects used, along with their affiliation and macro-area (Table 3/4).

ISO 639-3	name	family	macroarea
PPOWBT	Folopa	Teberan	Papunesia
PRKBSM	South Wa	Austroasiatic	Eurasia
PUIABC	Puinave	isolate	South America
QUBPBS	Huallaga Huánuco Quechua	Quechuan	South America
RAWBIB	Rawang	Sino-Tibetan	Eurasia
RELBTL	Rendille	Afro-Asiatic	Africa
ROOWBT	Rotokas	North Bougainville	Papunesia
SABWBT	Buglere	Chibchan	North America
SGWBSE	Sebat Bet Gurage	Afro-Asiatic	Africa
SHKBSS	Shilluk	Nilotic	Africa
SPPTBL	Supyire Senoufo	Atlantic-Congo	Africa
SRNBSS	Sranan Tongo	Indo-European	South America
SSDWBT	Siroi	Nuclear Trans New Guinea	Papunesia
SURIBS	Mwaghavul	Afro-Asiatic	Africa
SXNLAI	Sangir	Austronesian	Papunesia
TABIBT	Tabasaran	Nakh-Daghestanian	Eurasia
TACPBC	Western Tarahumara	Uto-Aztecan	North America
TBGWBT	North Tairora	Nuclear Trans New Guinea	Papunesia
TCATBL	Ticuna	Ticuna-Yuri	South America
TCCBST	Barabayiiga-Gisamjanga	Nilotic	Africa
TCSWYI	Torres Strait-Lockhart River Creole	Indo-European	Australia
TEETBL	Huehuetla Tepehua	Totonacan	North America
TEOBSU	Teso	Nilotic	Africa
TFRWBT	Teribe	Chibchan	North America
THATSV	Thai	Tai-Kadai	Eurasia
TIHBSM	Timugon Murut	Austronesian	Papunesia
TIKWYI	Tikar	Atlantic-Congo	Africa
TLJWBT	Talinga-Bwisi	Atlantic-Congo	Africa
TOPTBL	Papantla Totonac	Totonacan	North America
TPIPNG	Tok Pisin	Indo-European	Papunesia
TPTTBL	Tlachichilco Tepehua	Totonacan	North America
TQOTQO	Toaripi	Eleman	Papunesia
TRCWBT	Copala Triqui	Otomanguean	North America
TUFWYI	Central Tunebo	Chibchan	South America
URATBL	Urarina	isolate	South America
URBWBT	Urubú-Kaapor	Tupian	South America
VIELHG	Vietnamese	Austroasiatic	Eurasia
WBABIV	Warao	isolate	South America
WIMWYI	Wik-Mungkan	Pama-Nyungan	Australia
XALIBT	Oirad-Kalmyk-Darkhat	Mongolic-Khitan	Eurasia
XAVTBL	Xavánte	Nuclear-Macro-Je	South America
XSUMEV	Sanumá	Yanomamic	South America
YADTBL	Yagua	Peba-Yagua	South America
YLEWBT	Yele	isolate	Papunesia
YSSYYV	Yessan-Mayo	Sepik	Papunesia
YUJWBT	Karkar-Yuri	Pauwasi	Papunesia
YUZNTM	Yuracaré	isolate	South America
YVATBL	Yawa	Yawa-Saweru	Papunesia
ZNEZNE	Zande	Atlantic-Congo	Africa
ZPMTBL	Mixtepec Zapotec	Otomanguean	North America

Table 7: Overview of doculects used, along with their affiliation and macro-area (Table 4/4).

rank	v_s	P	N doculects	avg. -log p	N tokens	token coverage
1	write	say+write	110	122.27	196	0.92
2	answer	answer+say	100	inf	240	0.97
3	heal	heal+sick	100	39.65	65	0.81
4	scribe	law+scribe	91	282.37	66	1.00
5	forgive	sin+forgive	91	56.45	64	0.98
6	repent	sin+repent	83	88.26	57	1.00
7	vinegar	wine+vinegar	81	30.01	6	1.00
8	widow	widow+woman	80	56.93	27	0.96
9	raise	dead+raise	78	53.38	79	0.84
10	faith	believe+faith	77	inf	279	0.91
11	come	to+come	75	64.33	802	0.66
12	loaf	bread+loaf	75	42.05	27	1.00
13	prostitute	prostitute+woman	67	36.82	13	1.00
14	bread	eat+bread	65	39.65	73	0.85
15	cup	wine+cup	65	31.98	20	0.61
16	drink	wine+drink	63	31.06	61	0.61
17	prophet	write+prophet	60	47.11	125	0.70
18	read	read+write	60	36.69	28	0.88
19	knock	knock+door	60	33.37	9	1.00
20	smoke	smoke+fire	60	29.15	13	1.00
30	life	life+eternal	54	56.87	142	0.68
60	branch	tree+branch	40	29.66	18	0.90
90	silver	money+silver	32	25.06	17	0.81
120	language	language+word	27	49.97	40	1.00
150	endure	suffer+endure	23	28.48	30	0.73
180	barrack	house+soldier	20	31.32	6	1.00
210	milk	child+milk	18	21.75	4	0.80
240	n't	n't+not	16	19.94	54	0.22
270	naked	garment+naked	14	23.81	14	0.78
300	key	key+door	13	23.05	6	1.00
330	tithe	priest+tithe	12	18.70	6	0.67
360	roll	tomb+roll	11	17.41	6	0.46
390	tax	money+tax	9	40.09	22	0.81
420	hypocrite	hypocrite+good	8	56.52	20	0.67
450	wave	wave+water	8	20.06	6	0.67
480	gentle	gentle+peace	7	29.58	14	0.78
510	hospitality	receive+house	7	20.28	3	1.00
540	doctrine	teach+true	6	24.29	9	0.56
570	reconcile	peace+with	5	40.68	10	0.62
600	muzzle	bind+mouth	5	20.65	2	1.00
630	divide	divide+self	4	36.37	14	0.41
660	star	star+heaven	4	23.53	11	0.38
690	ring	finger+ring	4	19.53	2	0.67
720	married	marry+married	4	16.43	4	0.40
750	summer	new+summer	3	26.52	3	1.00
780	spring	spring+water	3	21.59	6	0.12
810	laugh	ridicule+laugh	3	20.05	2	1.00
840	slaughter	bring+kill	3	18.35	3	0.60
870	resist	resist+write	3	16.74	3	0.33

Table 8: Select output of Step 1 (top-20 extractions and every 30th extraction after)

Pair	Language	Predicted Strategy	Verdict	Aligned tokens	Gloss	Source
wife+woman	Basque	synlexified	correct	emazte+emazte	*emazte* ('wife')	de Rijk (2008, p. 961)
wife+woman	Bora	unlexified	unlexified	méwakye+méwakye	*méwá* ('wife')	Thiesen and Weber (2012, p. 25)
wife+woman	Mískito	circumlexified	incorrect	maî+maî	*maia* ('spouse')	Salamanca (1988, p. 228)
dead+die	Arabic	synlexified	correct	مَت+مَت	نَمُت ('we die')	Abu-Chacra (2007, p. 49)
dead+die	Chechen	synlexified	correct	вела+вела	*vella* ('died')	Awde and Galaev (2014, p. 57)
dead+die	Indonesian	circumlexified	incorrect	mat+i	*mati* ('die')	Sneddon et al. (2010, p. 75)
king+throne	Indonesian	synlexified	correct	takhta+takhta	*takhta* ('throne')	Stevens and Schmidgall-Tellings (2010, p. 987)
king+throne	Kamasau	circumlexified	correct	king+sia	*sia* ('chair')	Sanders and Sanders (1994, p. 61)
go+way	Abau	underspecified	correct	ley+	*ley* ('go')	Lock (2011, p. 25)
go+way	Bora	underspecified	correct	péé+	*pééhi* ('go')	Thiesen and Weber (2012, p. 50)
go+way	Italian	underspecified	correct	va+	*andare* ('go')	Maiden and Robustelli (2013, p. 222)
go+way	Tok Pisin	circumlexified	correct	go+i	*go*('go')	Dutton and Thomas (1994, p. 364)
go+out	Basque	synlexified	correct	ilki+ilki	*ilki* ('go out')	Aulestia (1989, p. 302)
go+out	Zande	underspecified	correct	ndu+	*ndu* ('go, walk')	Pasch (2007, 172,173)
go+out	Indonesian	underspecified	correct	pergi+	*pergi* ('go')	Sneddon et al. (2010, p. 165)
take+way	Ndyuka	underspecified	correct	teke+	*teke* ('take')	Huttar and Huttar (1994, p. 10)
take+way	Abun	underspecified	correct	nai+	*nai* ('took')	Berry and Berry (1999, p. 67, 56)
take+way	Kyaka Enga	unlexified	unlexified	la+	*nyii* ('take')	Draper and Draper (2002)
take+way	Jamaican English	circumlexified	correct	tek+we	*tek* ('take'), *we* ('away')	Bailey (1968, p. 378)
door+open	Kwoma	circumlexified	correct	nubureja+tagwa	*tagwa* ('open'), *nubereja* ('door')	Bowden (1997, p. 208, 157)
door+open	Kamsá	synlexified	uncertain	atëfjna+atëfjna	*bësasa* ('door')	O'Brien (2018, p.178)
door+open	Cora	unlexified	unlexified	antácuunyaraca+antácuunyaraca	*cuuna* ('open a door')	Casad (2012, p. 91)
door+open	Iban	circumlexified	correct	pintu+muka	*pintu* ('door'), *muka* ('open')	Omar (1969, p. 16, 228)
clean+un	Ndyuka	circumlexified	incorrect	takuu+takuu	*takuu* ('evil')	Huttar and Huttar (1994, p. 62)
clean+un	Javanese	synlexified	correct	jahat+jahat	*jahat* ('evil')	Wedhawati et al. (2001, p. 155)
clean+un	Cree	circumlexified	correct	kisîpêkisitêw+ekâ	*kisîpêkisitêw* a(wash), *ekâ*,('un-')	Waugh et al. (2025)
whole+world	Bauza	underspecified	correct	+bak	*bak* ('ground')	Briley (1997, p. 6)
whole+world	Yorno So	underspecified	uncertain	puu+	*puu* ('all')	Heath (2014, p. 280)
whole+world	Indonesian	underspecified	incorrect	seluruh+	*seluruh* ('whole'), *dunia* ('world')	Sneddon et al. (2010, p. 41, 56)
to+world	Bora	underspecified	incorrect	vú+	*vu* (goal marker)	Thiesen and Weber (2012, p. 156)
to+world	Ndyuka	underspecified	incorrect	+goontapu	*goontapu* ('world')	Huttar and Huttar (1994, p. 62)
to+world	Basque	underspecified	incorrect	ra+mundu	*-ra* (allative marker)	de Rijk (2008, p. 50)
from+go	Koyra Chiini Songhay	underspecified	incorrect	koy+	*hun* ('leave')	Heath (1999, p. 80)
from+go	Italian	circumlexified	correct	di+partì	*partì* ('depart'), *di* ('from')	Maiden and Robustelli (2013, p. 366)
from+go	Komi	circumlexified	correct	ысь+мун	*muniy* ('depart'), *-is* (ablative affix)	Avril (2006, p. 90, 242)
go+up	Rotokas	synlexified	correct	ipa+ipa	*ipa* ('ascended')	Firchow (1974, p. 40)
go+up	Komi	synlexified	correct	кыпӧдчы+кыпӧдчы	*kaniy* ('ascend')	Avril (2006, p. 216, 90)
go+up	Timugon Murut	underspecified	correct	minongoi+	*ongoi* ('go')	Prentice (1971, p. 47)
enter+in	Fulfulde	synlexified	correct	natt+natt	*nattay* ('enter')	McIntosh (1984, p. 125)
enter+in	Jamaican English	circumlexified	correct	go+iin	*go* ('go'), *iin* (in)	Bailey (1968, p. 227)
enter+in	Karkar	underspecified	correct	+mek	*mik* ('in')	Rigden (n.d., p. 112)
answer+say	Tok Pisin	circumlexified	correct	bek+tok	*tok* ('say'), *bek* ('back')	Dutton and Thomas (1994, p. 5)
answer+say	Iban	synlexified	correct	nyaut+nyaut	*naut* ('answer')	Omar (1969, p. 228)
fish+net	Chipaya	synlexified	uncertain	anś+chíś	*ch'iz* ('fish')	Cerrón-Palomino (2006, p. 194)
fish+net	Kyaka Enga	circumlexified	correct	oma+nyuu	*oma* ('fish'), *nyuu* ('bag')	Draper and Draper (2002, p. 229, 291)
blind+eye	South Wa	circumlexified	correct	dug+ngai	*ngai* ('eye'), *duk* ('blind')	Shorto (2013, p. 19)
blind+eye	Garifuna	synlexified	correct	marihin+marihin	*maríhin* ('not see')	Haurholm-Larsen (2016, p. 201)

Table 9: Aligned tokens to seed pairs that were manually inspected and given a verdict with dictionaries and grammars

PoS pair	least and most often modally synlexified (N languages per pair)
Adposition+Noun (N=461; 18%)	**bottom** = about+thing (1) accord+with (1) voice+with (1) before+man (1) book+of (1) of+rich (1) of+star (1) of+sign (1) country+of (1) demon+of (1) **top** = mountain+on (116) before+foot (80) in+peace (72) disciple+of (46) of+son (41) demon+in (27) of+woe (26) gold+of (18) in+world (17) city+in (14)
Adposition+Verb (N=460; 19%)	**bottom** = before+fall (1) before+go (1) before+set (1) beg+to (1) owe+to (1) over+throw (1) belong+to (1) believe+in (1) bring+up (1) bind+with (1) **top** = rise+up (107) get+up (104) down+fall (95) cry+out (85) before+defile (84) down+sit (81) stand+up (67) out+release (60) at+marvel (49) cut+off (42)
Noun+Verb (N=408; 37%)	**bottom** = understand+word (1) bear+tree (1) sit+throne (1) language+speak (1) say+woman (1) fear+speak (1) man+name (1) man+right (1) enter+place (1) eye+open (1) **top** = bread+eat (173) law+write (164) apostle+send (163) eat+food (161) steal+thief (154) prophet+write (153) glory+worship (151) joy+rejoice (150) bondservant+serve (150) fish+take (146)
Verb+Verb (N=245; 20%)	**bottom** = become+know (1) beg+say (1) bear+give (1) come+touch (1) cry+say (1) hear+let (1) lead+stray (1) command+say (3) go+set (3) look+see (3) **top** = deceive+lie (162) suffer+torment (146) persecute+suffer (120) know+understand (109) hear+marvel (108) greet+kiss (106) eat+reap (95) come+send (95) know+see (93) find+see (92)
Noun+Noun (N=198; 80%)	**bottom** = beast+thing (1) sin+thing (1) house+master (1) fruit+wine (1) gift+sacrifice (1) thing+work (1) man+woman (1) news+word (1) brother+mother (2) bread+piece (3) **top** = boat+ship (186) boat+sea (186) horse+soldier (184) money+stone (182) demon+devil (181) month+moon (181) bird+dove (176) guard+soldier (175) fire+light (174) cloak+garment (172)
Adjective+Noun (N=87; 67%)	**bottom** = day+first (1) day+many (1) new+wine (1) sharp+sword (3) blind+man (4) such+thing (4) body+whole (4) certain+man (6) great+multitude (7) many+people (8) **top** = blind+eye (179) blood+dead (178) famine+hungry (172) afraid+fear (167) dead+tomb (166) angry+wrath (164) eternal+life (164) money+poor (163) parable+word (152) garment+naked (149)
Affix+Verb (N=86; 89%)	**bottom** = ation+save (1) believe+ful (1) ent+excel (2) ent+hear (2) ant+know (2) ful+write (3) ion+suffer (3) ion+relate (4) ance+repent (4) dom+know (4) **top** = er+teach (142) ant+serve (105) er+pray (105) re+turn (105) beware+self (102) appoint+dis (100) care+ful (90) be-ed+love (90) ion+oppress (87) er+sin (82)
Proper Noun+Verb (N=84; 9%)	**bottom** = Christ+die (1) God+worship (1) Isaiah+say (1) Jesus+answer (1) Passover+eat (1) Paul+say (1) Peter+say (1) Peter+answer (2) **top** = Peter+answer (2) Jesus+answer (1) Christ+die (1) Isaiah+say (1) God+worship (1) Passover+eat (1) Paul+say (1) Peter+say (1)
AFX+Noun (N=82; 86%)	bottom = ual+woman (1) body+ion (1) ion+sin (1) ence+word (1) ent+word (1) ly+word (1) ness+thing (1) ness+sin (1) s+side (1) flesh+ly (2) top = et+trump (178) enemy+st (166) a-ed+shame (144) st+war (136) com+passion (111) author+ity (93) cy+prophet (88) out+side (83) age+bond (80) fool+ish (74)
Particle+Verb (N=79; 20%)	**bottom** = come+to (1) crow+not (1) destroy+to (1) enter+to (1) hear+not (1) heal+to (1) release+to (1) love+not (1) stumble+to (1) stand+to (1) **top** = to+want (25) lose+not (8) circumcise+not (4) teach+to (2) not+want (2) crow+not (1) hear+not (1) release+to (1) come+to (1) stumble+to (1)

Table 10: Examples of most and least synlexified granularized seed doculect word pairs per part of speech (PoS) pair. Numbers in parentheses in the first column represent the total number of pairs and the proportion of pairs for which at least one doculect dominantly synlexifies that pair ('% synlex'). PoS abbreviations are [n]oun, [a]djective, [v]erb, [p]reposition, affi[x], proper na[m]e, par[t]icle

pair	N doc.	frequency pair
rise + up	107	29
get + up	104	19
down + fall	95	83
down + sit	81	42
stand + up	67	21
cast + out	33	39
out + pour	29	21
depart + from	29	11
come + down	9	69
go + up	8	61
down + throw	4	15
down + go	3	29
enter + in	3	191
out + throw	3	22
get + in	3	10
on + stand	3	17
go + out	2	117
from + rise	2	31
on + sit	2	66
come + out	1	132
come + from	0	107
fall + on	0	37
fall + in	0	29
fall + from	0	12
enter + to	0	175
come + to	0	1173
come + on	0	88
come + up	0	31
depart + to	0	11
come + in	0	312
cast + in	0	52
cast + to	0	38
go + in	0	211
go + on	0	68
get + to	0	20
fall + to	0	42
from + go	0	109
in + stand	0	51
in + sit	0	75
in + rise	0	12
in + put	0	52
in + pour	0	11
go + to	0	657
on + put	0	40
on + pour	0	13
out + put	0	10
in + throw	0	31
put + to	0	59
pour + to	0	10
rise + to	0	25
sit + to	0	29
stand + to	0	17
throw + to	0	70

Table 11: (Caused) motion events, their cross-doculectal frequency of being dominantly synlexified (N doc.), and their corpus frequency.

pair	N doc.	frequency pair
promise + say	63	57
false + say	54	23
answer + say	42	279
say + thunder	27	11
ask + say	27	185
confess + say	13	16
say + speak	12	329
sin + speak	12	32
ar + say	9	10
say + write	6	266
lie + say	4	30
command + say	3	135
speak + still	2	10
say + word	2	452
prophet + speak	1	13
language + speak	1	20
fear + speak	1	10
among + say	1	10
Isaiah + say	1	11
say + to	1	1675
say + woman	1	16
Peter + say	1	48
cry + say	1	75
say + still	1	10
beg + say	1	14
Paul + say	1	23

Table 12: Speech events, their cross-doculectal frequency of being dominantly synlexified (N doc.), and their corpus frequency.

	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	-14.57443	0.55514	-26.254	< 2e-16	***
log.pair.freq	1.20256	0.14200	8.469	< 2e-16	***
pos.type=ADP+VERB	0.55582	0.34499	1.611	0.107154	
pos.type=NOUN+NOUN	9.23598	0.41556	22.225	< 2e-16	***
pos.type=NOUN+VERB	3.37687	0.36528	9.245	< 2e-16	***
pos.type=other	4.26841	0.31352	13.615	< 2e-16	***
pos.type=VERB+VERB	1.34335	0.41376	3.247	0.001168	**
macroarea=Australia	-1.06364	0.21781	-4.883	1.04e-06	***
macroarea=Eurasia	-0.02568	0.14583	-0.176	0.860197	
macroarea=North America	-0.51993	0.13806	-3.766	0.000166	***
macroarea=Papunesia	-0.31067	0.12234	-2.539	0.011101	*
macroarea=South America	-0.68934	0.13487	-5.111	3.20e-07	***
AIC	118693.8				
residual degrees of freedom	506471				

Table 13: Detailed results of the logistic regression predicting synlexification.

Construction-Based Reduction of Translationese for Low-Resource Languages: A Pilot Study on Bavarian

Peiqin Lin[*1,2], **Marion Thaler**[*1], **Daniela Goschala**[1], **Amir Hossein Kargaran**[1,2], **Yihong Liu**[1,2], **André F. T. Martins**[3,4,5], **Hinrich Schütze**[1,2]
[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning [3]Instituto Superior Técnico (Lisbon ELLIS Unit)
[4]Instituto de Telecomunicações [5]Unbabel
linpq@cis.lmu.de, Marion.Thaler@campus.lmu.de

Abstract

When translating into a low-resource language, a language model can have a tendency to produce translations that are close to the source (e.g., word-by-word translations) due to a lack of rich low-resource training data in pretraining. Thus, the output often is translationese that differs considerably from what native speakers would produce naturally. To remedy this, we synthetically create a training set in which the frequency of a construction unique to the low-resource language is artificially inflated. For the case of Bavarian, we show that, after training, the language model has learned the unique construction and that native speakers judge its output as more natural. Our pilot study suggests that construction-based mitigation of translationese is a promising approach. Code and artifacts are available at https://github.com/cisnlp/BayernGPT.

1 Introduction

The multilingual capabilities of large language models (LLMs) are impressive for medium- and high-resource languages, but they are still poor for low-resource languages for which the size of the available text corpus is small. While LLMs have recently improved their performance on low-resource comprehension tasks, little progress has been made on generation since the training demands for effective generation are much higher than those for comprehension. Bavarian is a low-resource language that instantiates this state of affair: some large state-of-the-art models' performance is decent for comprehension of Bavarian, but this does not carry over to generation.

Our hypothesis is that there are at least two different problems with limited generation capabilities of LLMs: lack of knowledge and translationese behavior.

Lack of knowledge mainly results in poor lexical choices. For example, our trained model (see below) translates German "Kuchen" 'cake' not as the correct Bavarian "Kuacha", but as "Kuchel" 'kitchen'. There is some promising work that addresses the lack of knowledge problem by prompting the LLM with relevant dictionary entries in in-context learning.

However, apart from the lack-of-knowledge problem, there is a second problem with the Bavarian generations of language models: translationese.

Translationese is a particular problem in machine translation with language models. The LMs tend to stick closely to the source sentence, especially when translating from a high-resource language to a closely related low-resource language as is the case for Standard German and Bavarian. Bavarian and Standard German are in a state of diglossia where Bavarian speakers produce forms of Bavarian that are closer to Standard German in more formal contexts and forms of Bavarian that can be completely incomprehensible to Standard German speakers in informal contexts.

This means that the Bavarian translationese generated by LMs is not necessarily incorrect: it may be appropriate Bavarian for certain contexts of language use. But clearly, the LMs do not have full competence of the Bavarian language if all they do is produce translationese.

In this paper, we take a small step towards addressing the translationese problem by training LMs to generate a Bavarian construction that does not occur in Standard German. This reduces the translationese property of what the LM generates because the output has clear indicators of being "genuine" Bavarian.

Specifically, we experiment with the article reduplication construction in Bavarian:

Bavarian	Ea woa friara a recht a fidel Buam.
Gloss	He was formerly a RD-modifier a jolly boy.
translation	He used to be quite a jolly boy.

[*]Equal contribution.

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 114–121
August 1, 2025 ©2025 Association for Computational Linguistics

With certain reduplication modifiers (RD modifiers), in particular with "recht", "so" and "ganz", this Bavarian construction consists of the reduplication of the indefinite article, with the RD modifier occurring between the two indefinite articles.

We show that a model trained with data synthetically generated to contain article reduplication learns to produce the construction, reducing the translationese character of the language model translations.

To summarize, our method translates an originally Bavarian corpus to German using a state-of-the-art LM, resulting in an "unmodified" parallel corpus; generates a "modified" parallel corpus by semi-automatically editing parallel sentences (such that the Bavarian sentence contains a Bavarian construction and the German sentence is modified to reflect that change) and trains LMs on modified and unmodified corpora. We also create two evaluation datasets, one for sentences, one for noun phrases. We manually evaluate the performance of the two trained models. We find that the model trained on modified data successfully produces article reduplication and its output data is perceived as less "translationese" than the generations of the model trained on unmodified data.

2 Related Work

Multilingual language models have emerged as the dominant paradigm for supporting low-resource languages. These models range from smaller architectures such as multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mBART (Liu et al., 2020), mT5 (Xue et al., 2021), and Glot500 (Imani et al., 2023), to large-scale models including BLOOM (Scao et al., 2022), Aya (Üstün et al., 2024), MaLA500 (Lin et al., 2024a), EMMA500 (Ji et al., 2024), and Llama 3 (Dubey et al., 2024). Trained jointly on data from a wide range of languages, these models demonstrate strong cross-lingual transfer and generalization capabilities, offering a promising foundation for low-resource language applications.

Despite this progress, generative performance on low-resource languages remains limited – particularly for tasks such as machine translation (MT), which are highly sensitive to the quantity and quality of available training data.

MT has thus become a central benchmark for evaluating the generative abilities of multilingual models. In the context of large language models, recent efforts have explored two major directions: supervised fine-tuning on parallel corpora (Yang et al., 2023; Xu et al., 2023, 2024; Lin et al., 2024b; Alves et al., 2024; Rei et al., 2024), and in-context learning methods that incorporate external linguistic resources – such as grammar books and bilingual dictionaries – without modifying model weights (Lu et al., 2023; Tanzer et al., 2024; Zhang et al., 2024b,a; Pei et al., 2025).

These challenges are particularly pronounced for extremely low-resource languages such as Bavarian. Due to very limited annotated data, Bavarian remains largely excluded from multilingual pretraining. Her and Kruschwitz (2024) introduced one of the first Bavarian–German MT systems, demonstrating that translation between closely related language varieties can yield relatively strong BLEU scores. To further enhance translation quality while minimizing artifacts such as translationese, they employed back-translation (Sennrich et al., 2016) to generate a compact but effective set of synthetic training examples. However, their approach depends solely on WikiMatrix (Schwenk et al., 2021), a parallel corpus known to be noisy and dominated by simplistic sentence structures, which limits its ability to robustly capture more nuanced translation characteristics.

One such characteristic is translationese – a linguistic phenomenon that arises when translated text retains unnatural or non-native structures. This artifact is especially problematic for low-resource languages (Graham et al., 2020). To address it, Chowdhury et al. (2022) proposed removing translationese signals implicitly encoded in vector embeddings, leading to improved performance on natural language inference tasks. Similarly, Wein and Schneider (2023) employed Abstract Meaning Representation (AMR) to abstract away surface-level features and suppress translationese. While effective, these techniques do not explicitly assess whether the resulting text resembles naturally written language. More recently, Jalota et al. (2023) evaluated the success of style transfer techniques in mitigating translationese by analyzing classifier performance before and after post-editing. Kunilovskaya et al. (2024) further explored the use of GPT-4 to mitigate translationese by incorporating linguistic cues into the prompting context. Complementarily, Kuwanto et al. (2024) introduced a storyboard-based data collection method, in which native speakers generate descriptions from visual prompts without access to the source text—resulting in

outputs that are more fluent and natural. However, these methods still fall short of enabling large language models to directly produce fluent, translationese-free output for truly low-resource languages, such as Bavarian.

3 Training

3.1 Data Preparation

Due to the scarcity of high-quality Bavarian–German parallel corpora, we use `GPT-4` to translate the Bavarian portion of the Wikipedia[1] into English and standard German. We use language identification (Kargaran et al., 2023) to filter out noise, such as when the model partially translates the source or directly copies it. From the resulting 22,564 Bavarian-English-German parallel documents, we reserve 1,000 for validation and another 1,000 for testing, with the remainder used for training. To create a sentence-level corpus, we segment the documents using line breaks and remove duplicate entries.

To reduce translationese effects and encourage native-sounding Bavarian output, we augment the original corpus using a rule-based algorithm grounded in syntactic analysis. We employ `spaCy` to parse the Standard German sentences and identify noun phrase structures of the form *indefinite article + adjective + noun*. These constructions serve as reliable anchors for inserting article reduplication in the aligned Bavarian sentence.

The algorithm first scans each tokenized German sentence for sequences where an indefinite article (e.g., *ein*, *eine*) is immediately followed by an adjective and a noun. To avoid semantically awkward or ungrammatical insertions, the algorithm filters out adjectives derived from nationalities (e.g., *deutsch*, *österreichisch*). For every such match, we check whether the corresponding Bavarian sentence has an equivalent syntactic pattern beginning with a Bavarian indefinite article (e.g., *a*, *oa*).

If the alignment is valid, we apply a transformation that inserts a reduplicated indefinite article separated by an RD modifier (randomly chosen from *recht*, *so*, *ganz*) between the original article and adjective. To maintain semantic alignment, the German counterpart is modified by inserting the intensifier *sehr* between the article and adjective.

This pipeline was run over sentence-aligned data and executed only where the token count matched between the Bavarian and German sides, ensuring high-precision transformations. Table 1 shows a representative example.

Before article reduplication transformation
Bavarian: *A heiliga Lebnsbaam* German: *Ein heiliger Lebensbaum*
After article reduplication transformation
Bavarian: *A recht a heiliga Lebnsbaam* German: *Ein sehr heiliger Lebensbaum*

Table 1: Example of article reduplication transformation in Bavarian–German parallel data.

3.2 Model Training

We develop a German-to-Bavarian machine translation system by instruction-tuning the `Llama 3.1 8B Chat` model (Dubey et al., 2024).

To accomplish this, we design a structured prompt format, as shown in Table 2. In this format, `[DEU_TEXT]` represents the input German sentence, and `[BAR_TEXT]` corresponds to the expected Bavarian translation. During training, both sentences are provided to the model, while at inference time, the model generates `[BAR_TEXT]` from the input German sentence.

To enable efficient fine-tuning, we use LoRA (Hu et al., 2022). The model is fine-tuned with a learning rate of 1×10^{-4}, weight decay set to 0.1, and the LoRA rank configured to 32.

We train two machine translation models:

- **m-base:** Trained on the original parallel dataset.
- **m-aug:** Trained on the dataset augmented with rule-based transformations.

4 Evaluation

To assess the effectiveness of our article reduplication augmentation strategy, we conducted both sentence-level and noun phrase–level (NP) evaluations using human judgments from two native Bavarian speakers (two of the authors of this paper).

4.1 Sentence-Level Evaluation

We used a test set of 141 Bavarian–Standard German sentence pairs which received the same augmentation as the training data of m-aug. Standard German inputs were translated into Bavarian by both m-base (baseline) and m-aug (augmented). The evaluation focused on three criteria:

[1] `dumps.wikimedia.org/barwiki`

```
<|start_header_id|>user<|end_header_id|>
Translate the following text from German to Bavarian.
German:  [DEU_TEXT]
Bavarian:  <|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
[BAR_TEXT]<|eot_id|>
```

Table 2: Prompt format used for instruction-tuned machine translation from German to Bavarian.

- **Correct application of article reduplication** – Is article reduplication used where grammatically appropriate?
- **Idiomatic and grammatical correctness** – Is the output of m-aug more idiomatic and grammatically natural and correct than m-base?
- **Pragmatic appropriateness** – Is article reduplication contextually suitable within the sentence?

The importance of pragmatic appropriateness can be illustrated by the following example.

> **Bavarian (with article reduplication):**
> *In da Eingobaauffordarung kennt ma an Untaschied zwischn Root und andan Nutzan duach* **a ganz a obschliaßends Rautzeichen** *(#) stott des Dollarzeichens ($).*
>
> **Standard German:**
> *In der Eingabeaufforderung erkennt man einen Unterschied zwischen Root und anderen Nutzern durch* ***ein sehr abschließendes Rautezeichen*** *(#) anstelle des Dollarzeichens ($).*
>
> **English translation:**
> *In the command prompt, one can recognize a difference between Root and other users by* ***a very final hash sign*** *(#) instead of the dollar sign ($).*

The emphatic use of article reduplication (*a ganz a obschliaßends Rautzeichen*) is unnatural and non-idiomatic in this technical context. As a result, it was evaluated as pragmatically inappropriate, even though the grammatical structure is correct.

The resulting outputs were evaluated by two native speakers. The evaluation regarding pragmatic appropriateness was conducted on the 70 sentences where reduplication was applied. The overall inter-annotator agreement was 100%, indicating high reliability of the judgments.

m-aug failed to apply article reduplication where grammatically possible in only 12 cases, and in just 19 cases the translation of m-base was assessed as more grammatically correct and idiomatic. However, regarding pragmatic appropriateness, there is a higher number of questionable cases. This is primarily due to the fact that the augmented training data was not filtered for pragmatic appropriateness, potentially including instances where article reduplication is not suitable. As such, m-aug provides a baseline that could be further improved with more appropriate training data.

Category	Count	Percentage
Reduplication correctly applied	70	49.65%
Not applicable (grammatically)	59	41.84%
Reduplication missed (applicable)	12	8.51%
Total	141	100.00%

Table 3: Sentence-level evaluation: Article reduplication accuracy.

Comparison Result	Count	Percentage
m-aug sentence is better	103	73.05%
Sentences are equivalent	19	13.48%
m-aug sentence is worse	19	13.48%
Total	141	100.00%

Table 4: Sentence-level evaluation: Idiomatic and grammatical correctness comparison (m-base vs m-aug).

Evaluation Result	Count	Percentage
Reduplication is pragmatically correct	42	60.00%
Reduplication is questionable	28	40.00%
Total	70	100.00%

Table 5: Sentence-level evaluation: Pragmatic appropriateness of reduplication.

4.2 Noun Phrase–Level Evaluation

Given that article reduplication targets noun phrases, we conducted a focused evaluation. A test set of 200 Standard German NPs in the structure *indefinite article* + *intensifier* + *adjective* + *noun* was generated using random combinations of *adjective* + *noun* from the Wikipedia corpus. The translations of both models were evaluated by a native speaker. This evaluation focused on whether the article reduplication was applied accurately and whether the idiomatic and grammatical correctness was improved compared to the NPs produced by the original Model A.

Category	Count	Percentage
Reduplication applied	200	100%
Reduplication not applied	0	0%
Total	200	100%

Table 6: NP-level evaluation: Article reduplication accuracy.

Comparison Result	Count	Percentage
NP of Model B is better	189	94.5%
NP of Model B is worse	11	5.5%
Total	200	100%

Table 7: NP-level evaluation: Idiomatic and grammatical correctness (Model A vs. B).

These results indicate that Model B systematically learned the reduplication pattern within the structure *indefinite article + intensifier + adjective + noun*, producing outputs that are both idiomatic and grammatically well formed.

To assess whether the model overgeneralizes article reduplication, we conducted a complementary evaluation using 200 Standard German noun phrases of the form *indefinite article + adjective + noun*, i.e., without an intensifier. This test aimed to determine if the model incorrectly applies reduplication to structures where it is not licensed. Only 4 out of 200 outputs contained article reduplication without an intensifier present. Interestingly, these instances were all triggered by the word *ganz* used adjectivally, as in the Standard German phrase *ein ganzer Ortsteil*, translated in Bavarian as *a ganza **a** Orstei* (gloss: *a whole subdistrict*). In these cases, *ganz*, previously encountered as an RD-modifier in the training data, was likely misinterpreted by the model as licensing reduplication, even when used adjectivally.

Category	Count	Percentage
Reduplication falsely applied	4	2%
Reduplication not applied	196	98%
Total	200	100%

Table 8: NP-level evaluation: Reduplication overgeneralization in NPs without intensifiers.

These findings suggest that the model applies article reduplication in a targeted and controlled manner, largely avoiding false positives.

5 Conclusion

We propose a method to remedy the problem of translationese when translating to low-resource languages and apply it to Bavarian. Our approach synthetically creates a training set in which the frequency of a construction unique to the low-resource language is artificially inflated. We show that a model trained with this synthetic data produces output with this construction and that it is perceived as being more natural than the baseline.

Limitations

Our pilot study has numerous limitations.

- We know of no linguistic studies that quantify the impact of constructions on the "naturalness" of linguistic production. Other factors may also have to be addressed to produce fully natural output.
- For a given pair of high resource and low resource languages, there may be no constructions that meet our selection criterion: that is, they are relatively frequent in the low-resource language and do not occur at all in the high-resource language.
- The construction we chose is easy to match and to generate. For more complex constructions, there is a risk that the modified low-resource sentences would not be correct, thereby introducing a new source of errors.
- We only implemented a baseline method for changing source and target languages. There are several ways this baseline method can be improved, e.g., by trying to eliminate the pragmatically inappropriate language we detected in the experiments.
- Our evaluation is very basic. Due to the difficulty of finding native speakers of Bavarian, two of the authors (who are native speakers of Bavarian) performed the annotation.

Acknowledgements

This work was funded by DFG (SCHU 2246/14-1), the European Research Council (DECOLLAGE, ERC-2022-CoG #101088763), EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the

Federal Ministry of Education and Research, and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, João Alves, M. Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef van Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3983–3991. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 72–81. Association for Computational Linguistics.

Wan-Hua Her and Udo Kruschwitz. 2024. Investigating neural machine translation for low-resource languages: Using bavarian as a case study. *CoRR*, abs/2404.08259.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1082–1117. Association for Computational Linguistics.

Rricha Jalota, Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7086–7100. Association for Computational Linguistics.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2024. EMMA-500: enhancing massively multilingual adaptation of large language models. *CoRR*, abs/2409.17892.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *Findings*

of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pages 6155–6218. Association for Computational Linguistics.

Maria Kunilovskaya, Koel Dutta Chowdhury, Heike Przybyl, Cristina España-Bonet, and Josef van Genabith. 2024. Mitigating translationese with GPT-4: strategies and performance. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), EAMT 2024, Sheffield, UK, June 24-27, 2024*, pages 411–430. European Association for Machine Translation (EAMT).

Garry Kuwanto, Eno-Abasi Urua, Priscilla Amondi Amuok, Shamsuddeen Hassan Muhammad, Aremu Anuoluwapo, Verrah Otiende, Loice Emma Nanyanga, Teresiah W. Nyoike, Aniefon D. Akpan, Nsima Ab Udouboh, Idongesit Udeme Archibong, Idara Effiong Moses, Ifeoluwatayo A. Ige, Benjamin Ajibade, Olumide Benjamin Awokoya, Idris Abdulmumin, Saminu Mohammad Aliyu, Ruqayya Nasir Iro, Ibrahim Said Ahmad, Deontae Smith, Praise-EL Michaels, David Ifeoluwa Adelani, Derry Tanti Wijaya, and Anietie Andy. 2024. Mitigating translationese in low-resource languages: The storyboard approach. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11349–11360. ELRA and ICCL.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024a. Mala-500: Massive language adaptation of large language models. *CoRR*, abs/2401.13303.

Peiqin Lin, André F. T. Martins, and Hinrich Schütze. 2024b. A recipe of parallel corpora exploitation for multilingual large language models. *CoRR*, abs/2407.00436.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *CoRR*, abs/2305.06575.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schütze. 2025. Understanding in-context machine translation for low-resource languages: A case study on manchu. *CoRR*, abs/2502.11862.

Ricardo Rei, José Pombal, Nuno Miguel Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tânia Vaz, Duarte M. Alves, M. Amin Farajian, Sweta Agrawal, António Farinhas, José Guilherme Camargo de Souza, and André F. T. Martins. 2024. Tower v2: Unbabel-ist 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 185–204. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1351–1361. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *CoRR*, abs/2402.07827.

Shira Wein and Nathan Schneider. 2023. Translationese reduction using abstract meaning representation. *CoRR*, abs/2304.11501.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *CoRR*, abs/2401.08417.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *CoRR*, abs/2305.18098.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching large language models an unseen language on the fly. *CoRR*, abs/2402.19167.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *CoRR*, abs/2402.18025.

High-Dimensional Interlingual Representations of Large Language Models

Bryan Wilie[†], **Samuel Cahyawijaya**[‡], **Junxian He**[†], **Pascale Fung**[†]
[†]Hong Kong University of Science and Technology [‡]Cohere
bwilie@connect.ust.hk

Abstract

Large language models (LLMs) trained on massive multilingual datasets hint at the formation of interlingual constructs–a shared region in the representation space. However, evidence regarding this phenomenon is mixed, leaving it unclear whether these models truly develop unified interlingual representations, or present a partially aligned constructs. We explore 31 diverse languages varying on their resource-levels, typologies, and geographical regions; and find that multilingual LLMs exhibit inconsistent cross-lingual alignments. To address this, we propose an interlingual representation framework identifying both the shared interlingual semantic region and fragmented components, existed due to representational limitations. We introduce Interlingual Local Overlap (ILO) score to quantify interlingual alignment by comparing the local neighborhood structures of high-dimensional representations. We utilize ILO to investigate the impact of single-language fine-tuning on the interlingual alignment in multilingual LLMs. Our results indicate that training exclusively on a single language disrupts the alignment in early layers, while freezing these layers preserves the alignment of interlingual representations, leading to improved cross-lingual generalization. These results validate our framework and metric[1] for evaluating interlingual representation, and further underscore that interlingual alignment is crucial for scalable multilingual learning.

1 Introduction

Interlingua, a universal language-neutral representation, is pivotal for cross-lingual generalization. Grounded in both linguistic theories and computational practice, this concept aims to treat languages equitably and capture universal semantic structures independent of any specific language (Richens, 1958; Vauquois, 1968; Schubert, 1989; Rayner et al., 2010a; Johnson et al., 2017). The advent of LLMs trained on extensive multilingual corpora suggests the potential of interlingual constructs naturally emerging without any explicit objectives (Conneau et al., 2020a; Chang et al., 2022; Moschella et al., 2023; Wendler et al., 2024). This is attributed to their ability to map representations from different languages into a shared multilingual representation space (Pires et al., 2019; Libovický et al., 2020; Conneau et al., 2020b; Muller et al., 2021; Zhao et al., 2024; Zeng et al., 2025).

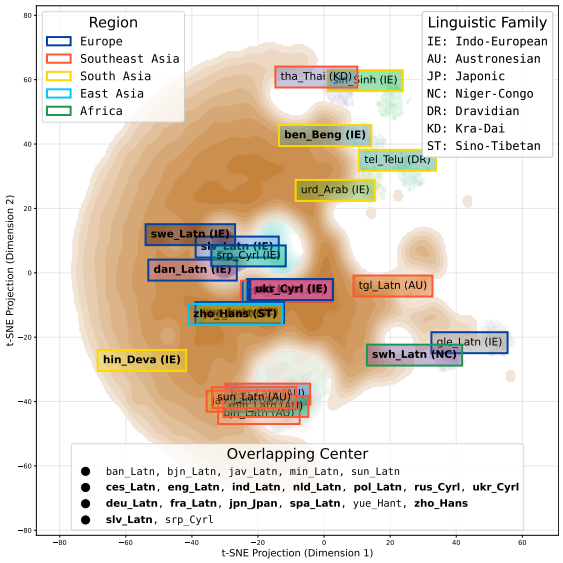


Figure 1: Interlingual overlaps transcending familial and regional boundaries in the intermediate layer, observed in a t-SNE visualization on the middle layer (16) of Aya Expanse (8B) hidden-state embeddings (HRLs in **bold**).

However, evidence remains mixed on whether they converge all language-specific representations into a unified single interlingual representation space, and raising questions about whether LLMs can retain the interlingual representations in diverse linguistic typology, geographical distribution, and resource-level settings. It is unclear whether LLMs form a unified interlingual construct or if fragmentation occurs across different language groups. A critical question persists: Do LLMs develop a uni-

[1] https://github.com/HLTCHKUST/interlingua

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 122–155
August 1, 2025 ©2025 Association for Computational Linguistics

versal interlingua representation, or present a partially aligned construct toward certain languages?

Our preliminary experiments reveal that LLMs represent parallel semantic input differently across languages. Notably, their neuron activations align better within high-resource pairs and the same familial or regional roots, suggesting that LLMs exhibit varying alignment consistencies across differing language groups. Building upon these insights, we introduce a novel interlingual representation framework aimed at enhancing the understanding of how LLMs encapsulate interlingual semantics. Our framework identifies both the core region that captures shared semantics across languages, and addresses fragmented components due to representational limitations underscoring the importance of interlingual alignment across diverse linguistic contexts. With the framework, we introduce a novel metric, Interlingual Local Overlap (ILO), which quantifies intrinsic interlingual alignment consistencies by comparing the local neighborhood structures of high-dimensional representations. Inspired by graph theory (Guimera and Amaral, 2005; Freeman et al., 2002; Borgatti and Everett, 2006), the ILO score is derived from the harmonic mean of two measurements, on the extent to which representations of a given language within the multilingual space: individually neighboring diverse other languages (**bridge**) and collectively connect diversely with other languages (**reachability**).

We demonstrate the effectiveness our framework and metric through an in-depth analysis of LLMs' internal states on a multilingual mathematical reasoning task, chosen for its language-agnostic properties. We first observe that training multilingual LLMs on a single-language causes catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999; Biesialska et al., 2020) degrading their cross-lingual generalization (Liu et al., 2021; Winata et al., 2023). These degradations are correlated with the disruption of interlingual alignment that originate in the early layers of LLMs. To ensure the preservation of interlingual alignments, we adopt a strategy of selectively-freezing parameters during the single-language fine-tuning. Evaluations using ILO highlight that this approach effectively safeguards the interlingual alignments across all layers and maintains the levels observed prior, which results in significant improvements in cross-lingual generalization. Ultimately, our findings underscore the pivotal role of interlingual semantic alignment in the pursuit of scalable multilingual learning.

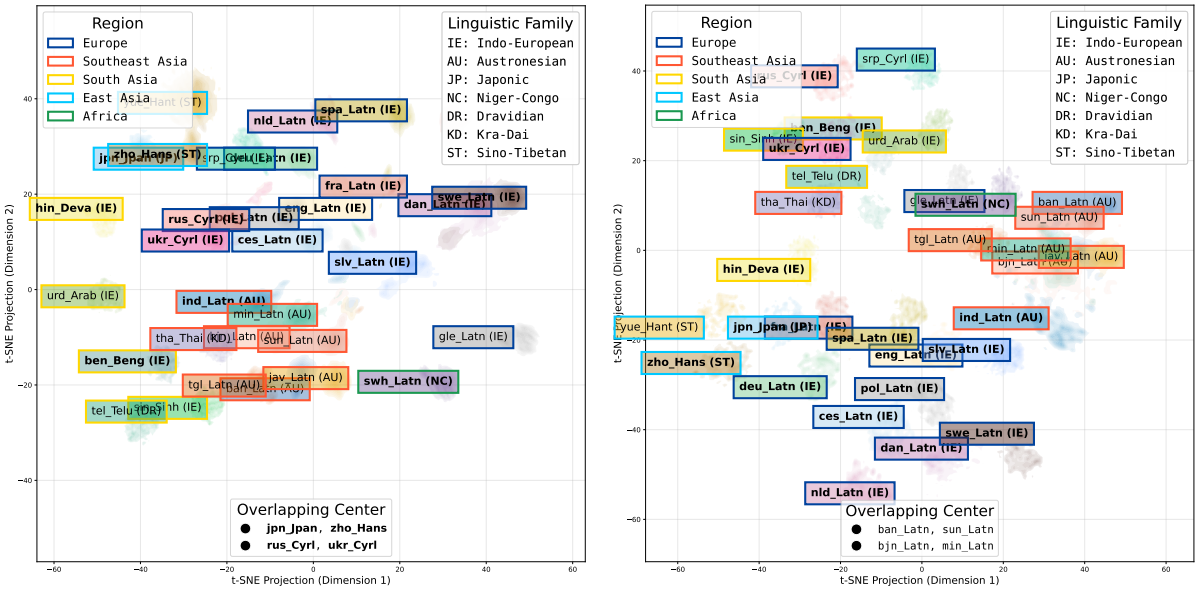
Properties	**Details**
Resources	High: 18 / Low: 13
Families	Indo-European: 18 / Austronesian: 7 / Sino-Tibetan: 2 / Japonic: 1 / Niger-Congo: 1 / Dravidian: 1 / Kra-Dai: 1
Regions	Europe: 14 / Southeast Asia: 8 / South Asia: 5 / East Asia: 3 / Africa: 1

Table 1: Distribution of the 31 languages across families, regions, and resource-levels in our analysis, sampled from Flores-200 (see Appendix A for complete details).

2 Related Works

Syntactical Interlingua Representations Interlingua has played a huge role throughout the development of NLP. Various representations of interlingua have been developed along with the advancement of NLP. In the early years, a logically formalized interlingua representation for mechanical translation has been proposed (Richens, 1958; Vauquois, 1968). In the early days, interlingua is presented as delexicalized grammar extracted from the original text that can be mapped to other language interlingua delexicalized grammar. In this case, each language has its own interlingua form which can then be mapped into other language with a dictionary lookup (Richens, 1958; Rayner et al., 2010b). A more sophisticated method involves interlingua representation as a common abstract syntax that are shared across all languages (Rayner et al., 2008; Kanzaki et al., 2008). This method has been applied in various systems such as Spoken Languge Translator (Rayner, 2000), PARC's XLE (Riezler et al., 2002), and Verbmobil (Wahlster, 2013). Despite its advancement, this method tends to be incomplete and difficult to scale to new languages (Ranta et al., 2020).

Semantic Interlingua Representations With the rise of statistical machine translation (Brown et al., 1990; Och et al., 1999; Lopez, 2008) and cross-lingual alignment (Brown et al., 1991; Och and Ney, 2003; Mikolov et al., 2013; Miceli Barone, 2016; Artetxe and Schwenk, 2019), methods for representing interlingua using latent semantic vectors become more prominent (Fung and Chen, 2004; Fung and Mckeown, 1994; Fung and Church, 1994; Seneff, 2006). Methods involving specialized objectives to construct better semantic interlingua representations have also been proposed (Lu et al., 2018; Al-Shedivat and Parikh, 2019; Zhu et al., 2020; Wei et al., 2021; Feng et al., 2022; Cahyawijaya et al., 2023, 2024b). In re-



(a) Early (Layer 0)

(b) Late (Layer 32)

Figure 2: Hidden-state embeddings of Aya Expanse (8B) projected in t-SNE dimensions (HRLs in **bold**). In these early and late layers, the representations cluster w.r.t resource levels and linguistic features, and minimally overlap.

cent years, various studies have showcased that current LLMs inherit such interlingua representation (Muller et al., 2021; Chang et al., 2022; Moschella et al., 2023; Zhao et al., 2024; Wendler et al., 2024) which enables LLMs to process sentences with a single shared representation across different languages. However, the characteristics of this representation in LLMs remain unexplored. This research aims to explore the extent of this interlingua representation offering a novel perspective on interlingual representation in LLMs.

3 Interlingual Representations in Multilingual LLMs

To explore the emergence of interlingual representation in LLMs, we assess the semantic alignment of their hidden states to understand whether the latent structures capture universal semantics across languages. We presume that multilingual LLMs adhere to a "first align, then predict" pattern (Muller et al., 2021) and that their aligned states represent semantically similar features across languages. Ideally, these features map parallel semantic inputs from many languages to similar vector representations that overlaps in the high-dimensional space.

Consider the high-dimensional representation space $\mathcal{H} \subseteq \mathbb{R}^d$ learned by LLMs, where d is the model's hidden-states dimension. For an input $\mathbf{x}$ in language ℓ, the model uses language-specific encoding functions $f_\ell(\mathbf{x}) \in \mathcal{H}$. Here, $\mathcal{H}$ serves as a shared multilingual space where different encoding functions $f_\ell(\mathbf{x})$ align semantic and syntactic patterns across languages. Building on this, we define semantic alignment α of representations from parallel inputs $\mathbf{x}$ and $\mathbf{x}'$ in languages ℓ and ℓ' as:

$$\alpha(\ell, \ell') = \mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim\mathcal{D}_{\ell,\ell'}} \left[\phi\left(f_\ell(\mathbf{x}), f_{\ell'}(\mathbf{x}')\right)\right].$$

Here, ϕ denotes a similarity function and $\mathcal{D}_{\ell,\ell'}$ is the distribution of semantically equivalent input pairs. A higher $\alpha(\ell, \ell')$ indicates better alignment.

3.1 Multilingual Shared Representation Space

We posit an interlingual representation framework that incorporates an intricate internal structure influenced by inherent model representational limitations. This framework highlights that the quality of alignment among representations may vary, leading to latent discrepancies that may stem from differences in resource availability or language-specific properties. Formally, we conceptualize the representations from various languages as falling into one of two qualitative regions of $\mathcal{H}$:

$$\mathcal{H} \supset \mathcal{M}_c \cup \bigcup_{\ell \in F} \mathcal{M}_{f_\ell}.$$

The component $\mathcal{M}_c$ is an aligned core interlingual region, that predominantly encapsulates shared semantics across languages. In contrast, the fragmented $\mathcal{M}_f$ represent regions where alignment with $\mathcal{M}_c$ is challenging. This framework refines the "first align, then predict" paradigm, that while LLMs align inputs from languages to a shared interlingual region, some remain partially aligned.

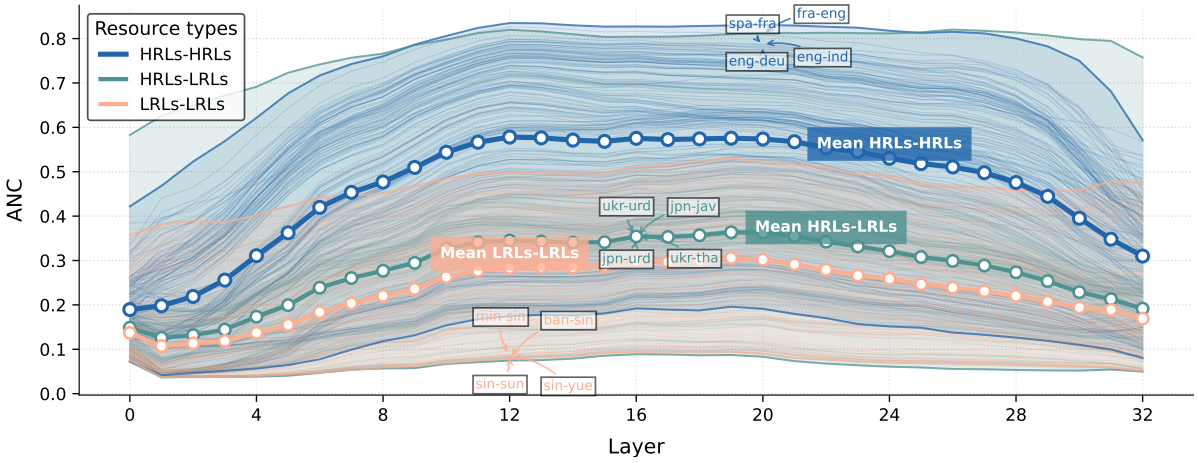


Figure 3: Comparisons of per-layer ANC scores on Aya Expanse (8B) with highlights on pairs w.r.t their resource-levels. Pairs of HRLs demonstrate strong correlations, while pairs involving LRLs exhibit lower ANC scores.

3.2 Core Interlingual Region

Conceptually, we define $\mathcal{M}_c$ as a region that predominantly encodes universal semantic structures and syntactic abstractions. By positioning multilingual representations in this shared region, LLMs effectively learn interlingual semantic representations that facilitate multilingual performance, e.g. through emphasizing semantics while minimizing language signals, retaining them only for language-specific predictions. This is where key interlingual alignments form, enabling LLMs to leverage universal semantic patterns for multilingual tasks.

3.3 Fragmented Region

While some languages enjoy substantial overlaps in $\mathcal{M}_c$, the less-aligned others occupy fragmented region $\mathcal{M}_f$ as they reflect model's representational limitation to embed the representation from these languages into $\mathcal{M}_c$. Factors such as sparse training data, typological distance, and morphological complexity might lead to partial alignment of these representations. Consequently, representations in $\mathcal{M}_f$ tend to be more weakly aligned to the universal semantics encoded by $\mathcal{M}_c$. This misalignment can degrade multilingual performances: tasks that rely on inputs from the less-aligned languages may exhibit lower performance since they draw from semantics that loosely intersects with $\mathcal{M}_c$.

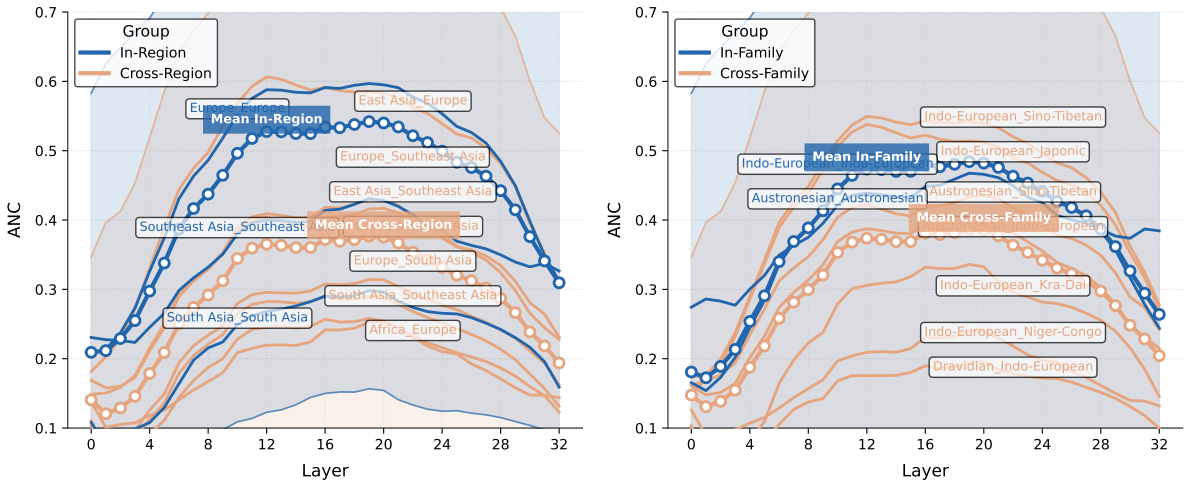
4 Semantic Alignment of Multilingual LLMs Representations

We explore the presence and characteristics of the components $\mathcal{M}_c$ and $\mathcal{M}_f$ within multilingual LLMs through assessing the semantic alignment between its hidden-states, derived from parallel inputs on various languages. Initially, we project LLMs' internal hidden-state embeddings into a 2D space to broadly assess proximities of parallel language representations and observe whether parallel input pairs in different languages clusters or overlaps. We then measure the cross-lingual alignment across the parallel hidden-state embeddings through neuron activation consistency w.r.t their resource-level, linguistic features, and geographical region.

We sample 31 diverse language subsets of Flores-200 (Team, 2024) varied on its resource-level, region, and family (Eberhard et al., 2024) (see Tables 1 and A1) as proxies to typological and morphological features (Georgi et al., 2010). Over experiments, we assess several multilingual LLMs: Aya Expanse (8B) (Dang et al., 2024), Llama-3.1 (8B) (Dubey et al., 2024), Gemma-2 (9B) (Team et al., 2024), Qwen-2.5 (7B) (Yang et al., 2024). We observe a universal phenomenon from these models, as described in the following sections. We put the further comparison details in Appendix D.

4.1 Inherent Regional Clustering with Mid-Layers High-Resource Alignment

We employ t-SNE (Van der Maaten and Hinton, 2008) to project LLMs' hidden-state embeddings into a 2D space and assess the proximities across language clusters. As t-SNE retains local neighborhood structures, overlaps in this 2D space imply closeness in the original high-dimensional space. In scenarios where representations are interlingually aligned, their nearest neighbors should comprise of multiple languages. We visualize the cross-lingual comparisons on the early, middle, and late layers of Aya Expanse (8B) in Figures 1 and 2, and others in Appendix C.2. We ran t-SNE with



(a) Highlights on pairs w.r.t their linguistic region

(b) Highlights on pairs w.r.t their linguistic family

Figure 4: Comparisons of per-layer ANC scores on Aya Expanse (8B) with highlights on pairs w.r.t their linguistic region and family. Consistently stronger alignments are observed between within-group mean correlations.

perplexity values of 5, 15, 30, 50, and observe consistent trends. Results for perplexity 15 are shown here; others are in Appendix F.1.

The t-SNE visualizations reveal distinct structural patterns across early (layer 0), intermediate (layer 16), and late (layer 32) layers (see Figures 2a, 1, 2b, respectively). In early and late layers, parallel language representations cluster according to resource levels and linguistic features, with minimal overlap. In contrast, the intermediate layer shows interlingual overlaps that transcend familial and regional boundaries, such as English and Russian overlapping with Indonesian, and Chinese with French. While overlaps mainly involve high-resource languages (HRLs), low-resource languages (LRLs) also exhibit overlaps, often due to regional factors. Meanwhile, some parallel representations remain fragmented outside these overlaps. These intermediate layer observations show that the quality of alignment varies. We further investigate the interactions in high-dimensional space to understand the alignment properties, in order to complement these low-dimensional observations.

4.2 Cross-lingual Alignments Depend on Resource-level and Linguistic Properties

Measurement. We further quantify the alignment characteristics by measuring neuron activation alignment for semantically identical inputs across different ℓ through *Average Neuron-wise Correlation* (ANC) (Del and Fishel, 2022). The ANC score in a certain LLM layer is defined as:

$$\text{ANC}(\ell, \ell') = \frac{1}{d} \sum_{i \in d} corr\left(f_\ell^i(\mathbf{x}), f_{\ell'}^i(\mathbf{x}')\right),$$

with $f_\ell^i(\mathbf{x})$ as the activation of i-th neuron for language ℓ and $corr$ denotes Pearson correlation between corresponding activations in ℓ and ℓ'. We visualize layer-wise ANCs from Aya Expanse in Figure 3 and 4, and others in Appendix B.

Findings. We find the "first align, then predict" patterns varies across language pairs. Notably, pairs of HRLs demonstrate strong correlations, while pairs involving LRLs exhibit lower scores (see Figure 3). Similarly, a consistent gap persists between within- and cross-group mean correlations, indicating stronger alignment within familial and regional language groups. Detailed analysis in Table A2 illustrates that most correlated pairs among LLMs are similar on their HRLs. Despite differing rankings, instruction-tuned LLMs exhibit similar sets of top language pairs with its pre-trained counterparts. These significant alignment gaps in cross-lingual correlations indicates latent discrepancies between semantically identical representations that stem from sparse data, typological distance, and the morphological complexity of languages.

5 Intrinsic Interlinguality of LLMs

In Section 4, we empirically demonstrated that multilingual LLMs' behavior aligns closely with the theoretical framework introduced in Section 3. Building upon these theoretical insights and empirical validations, we propose the Interlingual Local Overlap (ILO) score to measure the consistency of interlingual alignment in multilingual LLMs. Specifically, ILO score considers the local neighborhoods of models' hidden-state embeddings of

Dataset	Usage	# Lang	# Sample
Flores-200	Analysis	31	30,907
GSM8KInstruct	Training	10	73,559
MGSM	Evaluation	11	2,750

Table 2: Dataset statistics. "# Lang" indicates the number of languages represented in the dataset, and "# Sample" signifies the total sample size included.

linguistically-diverse semantically-parallel inputs, to indicate and quantify their intrinsic interlingual alignment in the high-dimensional space.

5.1 Interlingual Local Overlap Score

Given N input samples from set of languages in $\mathcal{L}$, $\{\mathbf{x}_i^\ell\}_{\ell\in\mathcal{L},i\in N}$, each sample $\mathbf{x}_i^\ell$ is embedded in model space $\mathcal{H}$ via $f_\ell(\mathbf{x})$. Let's denote $\mathcal{N}(\mathbf{x}_i^\ell)$ as the set of k-nearest neighboring languages of $\mathbf{x}_i^\ell$, defined as $\mathcal{N}(\mathbf{x}_i^\ell) = \{\ell' \neq \ell : \mathbf{x}_j^{\ell'} \in \mathrm{NN}_k(\mathbf{x}_i^\ell)\}$.

Bridge. The bridge score B_ℓ determines the degree of local interlingual mixing, analogous to the participation coefficient in graph theory, which assesses a node's link distribution across modules (Guimera and Amaral, 2005; Mijalkov et al., 2017). Bridge score measures the proportion of samples whose k-nearest neighbors in $\mathcal{H}$ include at least τ unique other languages, formally:

$$B_\ell = \frac{1}{N}\sum_{i\in N}\mathbf{1}\big(|\mathcal{N}(\mathbf{x}_i^\ell)| \geq \tau\big)$$

A score of ≈ 1 indicates that samples from ℓ consistently neighboring with diverse other languages.

Reachability. Inspired by classical degree of centrality in network analysis (Freeman et al., 2002; Borgatti and Everett, 2006), which quantifies a node's connections, we define reachability score to measure cross-lingual connectivity of ℓ representations. We view the multilingual space $\mathcal{H}$ as an undirected graph with each hidden-state embeddings as nodes linked to their k-nearest neighbors. The reachability score R_ℓ quantifies the connectivity degree of ℓ representations , defined as:

$$R_\ell = \frac{1}{|\mathcal{L}|-1}\left|\bigcup_{i\in N}\mathcal{N}(\mathbf{x}_i^\ell)\right|$$

R_ℓ enumerates the fraction of unique languages encountered across all samples of ℓ in $\mathcal{L}$, excluding itself. A high R_ℓ suggests that ℓ representations connect extensively within the multilingual space.

Interlingual Local Overlap (ILO). We then define an interlingual local overlap score ILO_ℓ to quantify the holistic interlingual alignment of language ℓ within $\mathcal{H}$, formally:

$$\mathrm{ILO}_\ell = 2\cdot\frac{B_\ell\cdot R_\ell}{B_\ell + R_\ell}$$

with the harmonic mean emphasizes the requirement of strong assessments in both the mixing and connectivity for the representations of ℓ to be considered as locally overlapping with other languages. Consequently, aggregated $\bar{\mathrm{ILO}}_\mathcal{L}$ of high ILO_ℓ in

$$\bar{\mathrm{ILO}}_\mathcal{L} = \frac{1}{|\mathcal{L}|}\sum_{\mathcal{L}}\mathrm{ILO}_\ell,$$

signals that multilingual LLMs effectively encode all of the diverse language inputs as aligned interlingual semantics within those in $\mathcal{L}$.

Preserving Interlinguality of LLMs. We demonstrate how ILO illuminate the performance variations in cross-lingual transfer and concurrently underscore the critical role of semantic interlingual alignment in multilingual LLMs. Cross-lingual transfer capitalizes on shared features to enhance multilingual capabilities (Philippy et al., 2023), typically involving single-language fine-tuning on a source language and directly applying it to target languages without further tuning. Despite its success, LLMs can suffer from catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999; Biesialska et al., 2020), where their cross-lingual generalization may degrade (Liu et al., 2021; Winata et al., 2023). Research suggests LLMs align multilingual inputs into language-independent representations, then revert them back to the query's original language (Muller et al., 2021; Zhao et al., 2024). Building on these insights, we conduct an experiment to preserve interlingual alignments by employing a **selective freezing** strategy, where we partially freeze parameters critical to language alignment. Our aim is to assess the potential mitigation of cross-lingual disruption, evaluated through ILO scores.

5.2 Experiment Design

To preserve the aligned semantics within multilingual model space, we experiment on freezing the parameters of the early layers on the first 4, 8, 12, and 16 layers. Additionally, we keep the token embedding, final layer normalization, and language modeling head (output projection layer) fixed. We identify these parameters as the language aligners.

Method	Training languages	Accuracy											Average	
		ben	tha*	swh	tel*	jpn	zho	deu	fra	rus	spa	eng	All	XL
Pre-trained	mixed	11.6%	12.0%	7.2%	0.0%	10.4%	8.8%	16.0%	12.4%	14.0%	11.6%	17.6%	10.3%	-
Fine-tuning	ben	**23.2%**	4.8%	1.2%	3.2%	10.0%	9.6%	10.8%	13.6%	11.6%	14.8%	12.8%	10.5%	9.2%
	tha*	1.6%	**32.8%**	4.4%	1.6%	14.4%	14.8%	17.2%	19.2%	18.0%	20.4%	25.6%	15.5%	13.7%
	swh	3.2%	6.4%	**30.8%**	2.8%	11.2%	12.4%	20.4%	19.6%	14.8%	22.4%	26.8%	15.5%	14.0%
	jpn	3.6%	7.2%	2.8%	1.2%	**32.8%**	21.6%	19.6%	18.0%	18.4%	22.4%	28.8%	16.0%	14.4%
	zho	0.8%	7.2%	2.4%	1.6%	22.0%	**34.8%**	19.6%	19.6%	21.6%	21.2%	27.6%	16.2%	14.4%
	deu	8.0%	16.4%	8.0%	**4.0%**	19.2%	19.6%	**37.6%**	**34.4%**	23.6%	28.8%	36.4%	21.5%	19.8%
	fra	4.8%	11.6%	4.0%	3.2%	16.0%	16.8%	31.6%	**34.4%**	25.6%	34.4%	35.6%	19.8%	18.4%
	rus	4.0%	14.0%	4.0%	1.2%	17.2%	16.4%	29.6%	28.4%	**34.0%**	30.0%	26.4%	18.7%	17.1%
	spa	4.8%	16.0%	2.8%	2.4%	14.4%	19.6%	28.4%	30.8%	31.2%	**38.4%**	38.4%	20.7%	18.9%
	eng	6.4%	14.4%	6.0%	2.4%	18.8%	24.4%	37.2%	27.2%	33.6%	33.2%	**43.2%**	**22.4%**	**20.4%**
Selective Freezing	ben	**26.4%**	12.8%	11.6%	14.4%	13.6%	14.8%	19.6%	20.0%	20.0%	17.6%	17.2%	17.1%	16.2%
	tha*	14.8%	**34.0%**	12.0%	12.4%	15.6%	21.6%	25.2%	22.0%	20.4%	24.4%	32.4%	21.3%	20.1%
	swh	9.2%	16.4%	**22.8%**	5.6%	14.0%	12.4%	18.4%	23.6%	19.2%	20.4%	27.6%	17.2%	16.7%
	jpn	16.0%	17.6%	12.0%	11.2%	**27.2%**	28.8%	24.4%	23.2%	24.0%	24.4%	29.6%	21.7%	21.1%
	zho	17.2%	17.2%	12.4%	12.0%	22.4%	**34.8%**	29.6%	22.4%	27.6%	23.6%	37.2%	23.3%	22.2%
	deu	12.8%	22.8%	14.4%	17.6%	20.0%	25.6%	**36.0%**	29.6%	27.6%	32.8%	39.2%	25.3%	24.2%
	fra	14.8%	24.8%	18.4%	12.0%	21.2%	21.2%	33.6%	**37.2%**	32.0%	**36.8%**	36.8%	26.3%	25.2%
	rus	20.4%	19.6%	11.6%	**18.8%**	22.0%	19.6%	28.8%	25.2%	38.4%	28.8%	32.0%	24.1%	22.7%
	spa	20.0%	24.0%	17.6%	16.8%	18.0%	27.2%	33.6%	33.6%	29.6%	34.0%	36.4%	26.4%	25.7%
	eng	20.4%	24.0%	18.0%	16.4%	20.4%	26.4%	35.2%	30.0%	**43.6%**	32.4%	**46.8%**	**28.5%**	**26.7%**

Table 3: Cross-lingual transfer performance on MGSM for Llama-3.1 (8B) without and with selective freezing. "XL" denotes average on languages that were not fine-tuned. Diagonal entries in blue highlights correspond to source language performances. Red highlights indicate decrease from pre-trained baseline. **Bold** and underline respectively denote the best within group and within column. The (*) marks languages classified as low-resource in Flores-200.

Datasets. We attend specifically to multilingual mathematical reasoning task, as it is inherently language-independent. We utilize the multilingual dataset `GSM8KInstruct` (Chen et al., 2024), which extends the English mathematical reasoning dataset `GSM8K` (Cobbe et al., 2021) by translating English instructions and chain-of-thought responses into 9 non-English languages via automatic translation and native-speaker human verification. To evaluate the model performance in this task, we utilize the `MGSM` benchmark (Shi et al., 2022). We attach the complete dataset statistics in Table 2.

Evaluation. We evaluate the accuracy of LLM greedy decoding zero-shot responses. Specifically, we employ the evaluation of Zhu et al. and determine answer accuracy by verifying that the final numerical value produced in the LLM's output exactly matches the ground-truth. In addition, we utilize ILO to investigate how changes in training impact LLMs' interlingual semantic alignment. To compute the ILO scores, we define a neighborhood size large enough to be informative and small enough to respect the local structures, while requiring each neighborhood to be rich in interlingual mixing. We experimented with Euclidean and cosine distance metric, with k, τ values of (5,3), (10,5), (20,10) and observe consistent trends. Results using $k = 10, \tau = 5$ and Euclidean distance are shown here; others in App. F.2. We evaluate the ILO scores using the same dataset from Section 4.

Models. We employ two multilingual LLMs: Llama-3.1 (8B) and Gemma-2 (9B). We train both LLMs using the same hyperparameters with learning rate $8e-5$, batch size 8, and gradient accumulation of 16 for 3 epochs using 4 A800 GPUs.

5.3 Results and Analysis

Cross-Lingual Transfer. We present findings from our cross-lingual transfer experiments, detailed in the Tables 3 and A3 within the "**fine-tuning**" rows, where we evaluated the performance of the fine-tuned Llama-3.1 and Gemma-2 respectively. Consistent with the expectations, we observed substantial cross-lingual transfer signified by improved performance in both source and target languages, even without direct training in those languages. The transfer is notably more pronounced in HRLs and languages within the same families and regions, such as the Indo-European languages in Europe: English, Spanish, Russian, French, and German. Remarkably, in some instances, performances on the target languages paralleled the ac-

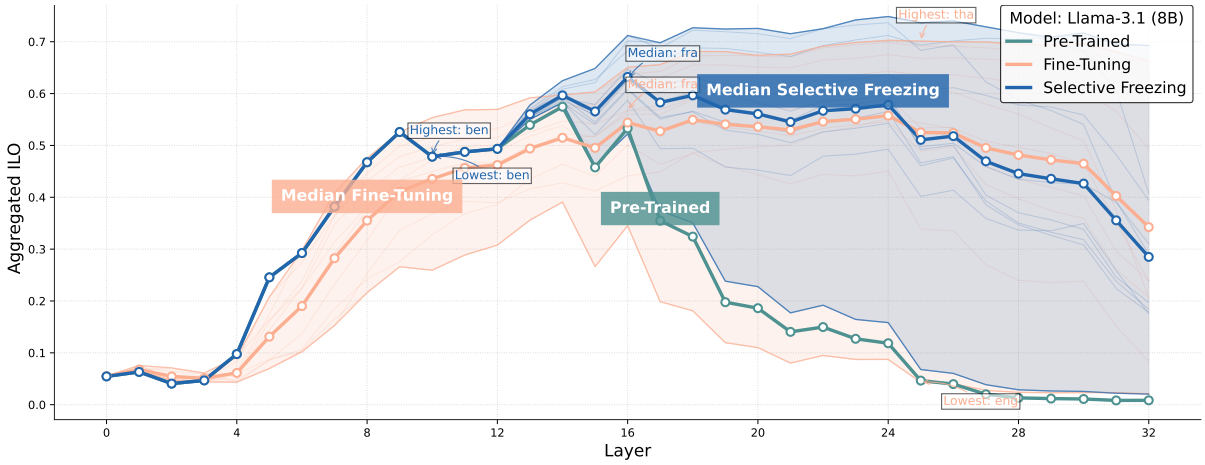


Figure 5: Layer-wise $\overline{\text{ILO}}_{\mathcal{L}}$ scores for all of the source-languages in the single-language training on Llama-3.1 (8B) in **pre-trained**, **fine-tuning**, and **selective freezing** modes. Decrease in alignment from single-language **fine-tuning** is seen in the early layers, whereas **selective freezing** allows LLM to sustain its **pre-trained** semantic alignment.

curacies in the source language, as exemplified by Spanish-to-English achieving 38.4%, which is on par with the Spanish-to-Spanish performance.

Despite the transfer, performance degradations are also observed on some of target languages. We conjecture that this issue stems from disruptions in the functionality of the aligner module. To investigate this hypothesis, we compute per-layer aggregated $\overline{\text{ILO}}_{\mathcal{L}}$ scores, and visualize them in Figures 5 and A1, for all of the source-languages trained on each the Llama-3.1 (8B) and Gemma-2 (9B) models. Both figures show a notable decrease in interlingual semantic alignment post fine-tuning that appears as early as in the 4^{th} layer for Llama and the 6^{th} layer for Gemma. Critically, the degree of alignment does not recover to the height of its pre-trained levels even after additional computational stages in subsequent layers. Furthermore, the interlingual overlaps initially present in the pretrained models become disrupted following single-language fine-tuning, as evidenced by reduced overlapping centers and loosened language clusters (Figs (b) of A14 vs A10, and A16 vs A12).

Preservation of LLMs' Interlinguality. Here we analyze the impact on freezing the first 12 layers, since it provides the best aggregated improvements (see Appendix F.3 for details). The quantitative analysis through the lens of the aggregated $\overline{\text{ILO}}_{\mathcal{L}}$ reveals that multilingual LLMs trained with **selective-freezing** mechanism sustain their prior semantic alignment levels in the early layers, and across all layers, as demonstrated in Figures 5 and A1. Empirical findings in Tables 3 and A3 further corroborate these insights, highlighting the substantial impact of maintaining interlingual semantic alignments on enhancing multilingual performances. Through keeping the aligner parameters unchanged, both LLMs understudy gain improved cross-lingual generalization compared to their post fine-tuning performances on source languages. Enhanced transfers can be observed on languages within-families and within-regions, with improvements and nearly no degradation towards the low-resource, cross-family, and cross-regional languages. Additionally, models fine-tuned with selective freezing effectively retain their original interlingual alignment, with overlapping centers largely preserved and clusters remaining tight (see Figs (b) of A15 vs A10, A17 vs A12, and App. E). These findings indicate that preserving the interlingual alignment in LLMs is essential for scalable multilingual learning. They emphasize the critical role of interlingual representation alignments in enhancing the multilingual capabilities of LLMs.

6 Conclusion

The emergence of multilingual LLMs demonstrates that interlingual constructs naturally arise, even in the absence of explicit objectives. We introduce a conceptual framework to understand interlingual representations, identifying both the core interlingual region that captures shared semantics, and fragmented components that reveal representational limitations in aligning with this core region. To advance the understanding of interlingual semantic alignment, we propose the Interlingual Local Overlap (ILO) score which quantifies alignment in the local neighborhood structures of interlingual

high-dimensional representations. Our proposed framework and metric illuminates the critical role of semantic alignment, offering a quantitative view into the high-dimensional alignment of multilingual representations. This study emphasizes interlingual semantic alignment and provides critical insights to optimize multilingual LLMs in the context of diverse linguistic tasks.

Limitations

Bias on linguistic family. In our analysis of interlingual regions, we sample 31 diverse languages from the Flores-200 set, representing various resource levels, geographical regions, and language families. We note, however, that there is a predominance of Indo-European languages within our HRLs subset. This distribution reflects the broader availability of linguistic data, as evidenced by web crawl statistics from CommonCrawl, where Indo-European languages are disproportionately represented. This imbalance is not intentional but rather an inherent limitation arising from existing data availability. Consequently, the observed stronger correlations among HRLs may partially reflect this underlying bias. We encourage future works to account for this, since observed correlations among HRLs may partially reflect this underlying bias.

Broader multilingual evaluations. Additionally, our study of cross-lingual transfer primarily utilizes multilingual mathematical reasoning task due to their largely language-agnostic nature. Such task allow us to simultaneously asses the linguistic understanding and logical reasoning capabilities of multilingual LLMs. We argue that the cross-lingual transfer capabilities evaluated within this work offer significant insights into general multilingual performance. Nonetheless, we encourage future studies to broaden evaluations to other tasks to extend the insights into interlingual alignment.

Expanding the core interlingual region. Our works presumes the existence of the core interlingual region where semantically aligned representations shared across languages, and others that only partially aligned to this core. Future works could explore on expanding this core interlingual region to encompass a broader range of languages, i.e. to introduce learning techniques that explicitly encourage deeper and more diverse interlingual mixing. Incorporating a larger, more heterogeneous multilingual datasets and leveraging linguistic priors might further strengthen the core region, and in turn, enhancing the universality of the core interlingual representations.

Bridging fragmented regions. A significant limitation of existing multilingual LLMs is that certain languages, particularly the underrepresented or typologically distant ones, most likely form fragmented region rather than being integrated fully with the core cluster. To address this, future work could aim to develop targeted strategies to encourage the integration of these regions and to narrow these gaps, i.e. under conditions of extremely limited data. Such interventions could facilitate the alignments of interlingual representation, thereby improving overall inclusivity and richness in linguistic diversity of the multilingual LLMs.

Predicting cross-lingual transfer. Although our work provides valuable insights into the local alignment of multilingual embeddings, it does not predict downstream cross-lingual transfer performance. One key limitation, for example, is that our proposals captures generic interlingual mixing of hidden-states representations and not the alignments of task vectors (Ilharco et al., 2022) that might be integral for effective transfer. This disconnect may arise when models achieve strong interlingual alignment while simultaneously losing critical nuances required for task performance. Future work could explore the integration of our proposals with task-aware signals, to develop quantifiers that are more designed to predict cross-lingual transfer.

Towards pure semantic representations. While our current work focuses solely on textual embeddings, a major frontier for future research lies in extending the framework of quantifying alignment via the local neighborhood structures of high-dimensional representations, to multimodal settings. Considering information from another modalities, it may be beneficial to disentangle and measure pure semantic content from modality-specific biases effectively. Exploring this direction not only hints promises to elucidate and improve modality-transfer but also potentially advance our understanding of how different forms of information interact to shape a universal semantic space. We envision our work, upon many others (e.g. Cahyawijaya et al. (2024a); Engels et al. (2025); Ji et al. (2024); Liu et al. (2024); Grosse et al. (2023)), to foster explorations towards the study of LLMs' semantic space.

Ethical Considerations

The exploration of interlingual representation in multilingual LLMs presents a unique opportunity to foster diversity and inclusivity in the field of NLP. Our work introduces framework and metrics to inspect interlingual representations in multilingual LLMs. They enable the analysis of interlingual alignment of different languages in the naturally emerging interlingual constructs within LLMs. We use publicly available parallel corpora and adhere to best practices in data handling, ensuring that no sensitive or personally identifiable information is involved. While our proposals help reveal disparities in representation, through this work, we instead leverage these insights to drive proactive interventions—ensuring future multilingual LLMs are not only more inclusive but also more reflective of the rich linguistic diversity they aim to serve. We hope our results contributes to more equitable model development and encourages further investigation into mitigating potential representational gaps across underrepresented languages.

Embracing Language Diversity Our work aims to create a universal representation that respects and preserves the unique characteristics of each language. Our findings highlight the importance of consistent interlingual alignments. By recognizing and capturing shared semantic structures through interlingua representations, LLMs can contribute to the preservation of linguistic diversity, ensuring that no single language or language group dominates the representation space. We envision LLMs to effectively represent and understand diverse languages, to be truly inclusive in language technology (e.g. Cahyawijaya (2024)). This is particularly crucial for underrepresented languages and communities, enabling them to have their voices heard and enabling them equal access of information, for example to their language-agnostic applications.

Addressing Bias and Fairness The study's observation of varying alignment consistencies across language groups underscores the need for careful consideration of bias. By identifying and addressing fragmented components due to representational limitations, we can work towards creating fairer representations. This is essential to prevent the reinforcement of existing biases and ensure equitable treatment of all languages. When LLMs effectively bridge the gap between languages, they enable seamless communication and understanding, benefiting diverse communities and fostering a more inclusive digital information systems.

References

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1184–1197.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541.

Stephen P Borgatti and Martin G Everett. 2006. A graph-theoretic perspective on centrality. *Social networks*, 28(4):466–484.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176.

Samuel Cahyawijaya. 2024. *Llm for everyone: Representing the underrepresented in large language models*. Ph.D. thesis, Hong Kong University of Science and Technology (Hong Kong).

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2024a. High-dimension human value representation in large language models. *arXiv preprint arXiv:2404.07900*.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024b. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. Instructalign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings*

of the First Workshop in South East Asian Language Processing, pages 55–78.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Maksym Del and Mark Fishel. 2022. Cross-lingual similarity of multilingual representations revisited. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 185–195.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are linear. In *The Thirteenth International Conference on Learning Representations*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Linton C Freeman et al. 2002. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge*, 1:238–263.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Pascale Fung and Benfeng Chen. 2004. Biframenet: bilingual frame semantics resource construction by cross-lingual induction. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 931–937.

Pascale Fung and Kenneth Ward Church. 1994. K-vec: A new approach for aligning parallel texts. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Pascale Fung and Kathleen Mckeown. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pages 385–393.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Roger Guimera and Luís A Nunes Amaral. 2005. Cartography of complex networks: modules and universal roles. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(02):P02001.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. Llm internal states reveal hallucination risk

faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kyoko Kanzaki, Yukie Nakao, Manny Rayner, Marianne Santaholma, Marianne Starlander, and Nikos Tsourakis. 2008. Many-to-many multilingual medical speech translation on a PDA. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Government and Commercial Uses of MT*, Waikiki, USA. Association for Machine Translation in the Americas.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. 2024. On the universal truthfulness hyperplane inside llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18199–18224.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71.

Adam Lopez. 2008. Statistical machine translation. *ACM Comput. Surv.*, 40(3).

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. *arXiv preprint arXiv:1804.08198*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany. Association for Computational Linguistics.

Mite Mijalkov, Ehsan Kakaei, Joana B Pereira, Eric Westman, Giovanni Volpe, and Alzheimer's Disease Neuroimaging Initiative. 2017. Braph: a graph theory software for the analysis of brain connectivity. *PloS one*, 12(8):e0178798.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2023. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina. 2020. Abstract syntax as interlingua: Scaling up the grammatical framework from controlled languages to robust pipelines. *Computational Linguistics*, 46(2):425–486.

Manny Rayner. 2000. *The spoken language translator*. Cambridge University Press.

Manny Rayner, Pierrette Bouillon, Beth Ann Hockey, and Yukie Nakao. 2008. Almost flat functional semantics for speech translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 713–720, Manchester, UK. Coling 2008 Organizing Committee.

Manny Rayner, Paula Estrella, and Pierrette Bouillon. 2010a. A bootstrapped interlingua-based smt architecture. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*.

Manny Rayner, Paula Estrella, and Pierrette Bouillon. 2010b. A bootstrapped interlingua-based SMT architecture. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

R. H. Richens. 1958. Interlingual machine translation. *The Computer Journal*, 1(3):144–147.

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell III, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Klaus Schubert. 1989. Interlinguistics–its aims, its achievements, and its place in language science. *Interlinguistics: Aspects of the Science of Planned Languages. Trends in Linguistics*, 42:7–44.

Stephanie Seneff. 2006. Combining interlingua with SMT. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Panel on hybrid machine translation: why and how?*, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, pages 1–6.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Bernard Vauquois. 1968. A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *IFIP Congress*.

Wolfgang Wahlster. 2013. *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *International Conference on Learning Representations*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Genta Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. 2023. Overcoming catastrophic forgetting in massively multilingual continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 768–777.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*.

Appendix

A Details on Linguistic Properties

We provide the detail of the region and linguistic properties of the language subsets sampled from Flores-200 in A1. Here, while most of them are extracted from Team (2024), we refer to (Eberhard et al., 2024) for the details on linguistic families.

Code	Language	Script	Region	Family	Res.
ban_Latn	Balinese	Latin	Southeast Asia	Austronesian	Low
ben_Beng	Bengali	Bengali	South Asia	Indo-European	High
bjn_Latn	Banjar	Latin	Southeast Asia	Austronesian	Low
ces_Latn	Czech	Latin	Europe	Indo-European	High
dan_Latn	Danish	Latin	Europe	Indo-European	High
deu_Latn	German	Latin	Europe	Indo-European	High
eng_Latn	English	Latin	Europe	Indo-European	High
fra_Latn	French	Latin	Europe	Indo-European	High
gle_Latn	Irish	Latin	Europe	Indo-European	Low
hin_Deva	Hindi	Devanagari	South Asia	Indo-European	High
ind_Latn	Indonesian	Latin	Southeast Asia	Austronesian	High
jav_Latn	Javanese	Latin	Southeast Asia	Austronesian	Low
jpn_Jpan	Japanese	Japanese	East Asia	Japonic	High
min_Latn	Minangkabau	Latin	Southeast Asia	Austronesian	Low
nld_Latn	Dutch	Latin	Europe	Indo-European	High
pol_Latn	Polish	Latin	Europe	Indo-European	High
rus_Cyrl	Russian	Cyrillic	Europe	Indo-European	High
sin_Sinh	Sinhala	Sinhala	South Asia	Indo-European	Low
slv_Latn	Slovenian	Latin	Europe	Indo-European	High
spa_Latn	Spanish	Latin	Europe	Indo-European	High
srp_Cyrl	Serbian	Cyrillic	Europe	Indo-European	Low
sun_Latn	Sundanese	Latin	Southeast Asia	Austronesian	Low
swe_Latn	Swedish	Latin	Europe	Indo-European	High
swh_Latn	Swahili	Latin	Africa	Niger-Congo	High
tel_Telu	Telugu	Telugu	South Asia	Dravidian	Low
tgl_Latn	Tagalog	Latin	Southeast Asia	Austronesian	Low
tha_Thai	Thai	Thai	Southeast Asia	Kra-Dai	Low
ukr_Cyrl	Ukrainian	Cyrillic	Europe	Indo-European	High
urd_Arab	Urdu	Arabic	South Asia	Indo-European	Low
yue_Hant	Yue Chinese	Han (Traditional)	East Asia	Sino-Tibetan	Low
zho_Hans	Chinese (Simplified)	Han (Simplified)	East Asia	Sino-Tibetan	High

Table A1: Complete distribution of the 31 languages across families, regions, and resource-levels in our analysis, sampled from Flores-200

B Further Details on ANC Scores

Here we provide a detailed view on the ANC comparison of the language pairs for all the model understudy. We compute aggregate peak score for each language pair as the mean over the peak layers. We identify the peak layer by computing the 75^{th} percentile of ANCs for each layer and select the top 3 layers as the peak layers. We denote all the top correlated language pairs from the layers with peak ANC scores and the unique languages from the top language pairs in Table A2. We find that the top correlated pairs with high ANCs among the LLMs are similar on their HRLs. Instruction-tuned LLMs exhibit similar sets of top language pairs with its pre-trained counterparts, despite the differing rankings of them.

C Visualization and Comparisons For Other Multilingual LLMs

C.1 ANC Comparisons from Other LLMs

We attach the complete visualization on ANC scores derived from the hidden-state embeddings of Aya Expanse (8B), Llama-3.1 (8B), Llama-3.1-Instruct (8B), Gemma-2 (9B), Gemma-2-Instruct (9B), and Qwen (9B), respectively in Figures A2, A3, A4, A5, A6, and A7.

C.2 T-SNE Visualizations from Other LLMs

We attach the complete t-SNE visualization projected from the hidden-state embeddings of Aya Expanse (8B), Qwen (9B), Llama-3.1 (8B), Llama-3.1-Instruct (8B), Gemma-2 (9B), and Gemma-2-Instruct (9B), respectively in Figures A8, A9, A10 A11, A12, and A13.

C.3 Reports on Cross-Lingual Transfer Experiments for Gemma-2 (9B)

We attach the cross-lingual transfer performance on MGSM and the layer-wise $\bar{\text{ILO}}_{\mathcal{L}}$ scores, for Gemma-2 (9B) in its pre-trained, fine-tuning, and selective-freezing modes, in Table A3 and Figure A1.

D Interlingual alignments of various multilingual LLMs

In this work, we observe a universal phenomenon that various multilingual LLMs, irrespective of their specific architecture or training data, exhibit a common behavior in constructing an interlingual representation region within their middle layers. However, amongst these similar general trend, we observe that there are different alignment levels across different LLMs in App B, C.1, and C.2)

For example, the t-SNE visualization of LLMs intermediate layers in Figures A12 and A8 shows that Gemma-2 (9B) exhibits more overlapping and closer clustering of language centers compared to Aya Expanse (8B). This observation is further supported by our neuron-wise correlation analysis, showcased in Figures A5 and A2, where the intermediate layers of Gemma-2 consistently show mean cross-lingual correlations exceeding 0.5, whereas in the intermediate layers of Aya Expanse, only the mean HRLs-HRLs and in-region records the correlations above 0.5. We conjecture that these variatons on alignment levels stem from the differences in the model architecture and training details of the LLMs.

Models	Gemma-2 (9B)	Gemma-2 It (8B)	Aya Expanse (8B)	Llama-3.1 (8B)	Llama-3.1 It (8B)	Qwen-2.5 (7B)
Top language pairs	dan_Latn - swe_Latn	eng_Latn - fra_Latn	rus_Cyrl - ukr_Cyrl	yue_Hant - zho_Hans	yue_Hant - zho_Hans	yue_Hant - zho_Hans
	eng_Latn - fra_Latn	dan_Latn - swe_Latn	eng_Latn - fra_Latn	rus_Cyrl - ukr_Cyrl	rus_Cyrl - ukr_Cyrl	dan_Latn - swe_Latn
	rus_Cyrl - ukr_Cyrl	rus_Cyrl - ukr_Cyrl	yue_Hant - zho_Hans	dan_Latn - swe_Latn	dan_Latn - swe_Latn	rus_Cyrl - ukr_Cyrl
	yue_Hant - zho_Hans	deu_Latn - eng_Latn	eng_Latn - ind_Latn	eng_Latn - fra_Latn	eng_Latn - fra_Latn	fra_Latn - spa_Latn
	dan_Latn - eng_Latn	yue_Hant - zho_Hans	fra_Latn - spa_Latn	fra_Latn - spa_Latn	fra_Latn - spa_Latn	eng_Latn - fra_Latn
	eng_Latn - swe_Latn	eng_Latn - swe_Latn	deu_Latn - eng_Latn	deu_Latn - swe_Latn	deu_Latn - swe_Latn	fra_Latn - rus_Cyrl
	deu_Latn - eng_Latn	dan_Latn - eng_Latn	ces_Latn - rus_Cyrl	deu_Latn - fra_Latn	deu_Latn - fra_Latn	rus_Cyrl - spa_Latn
	deu_Latn - swe_Latn	deu_Latn - fra_Latn	ces_Latn - ukr_Cyrl	deu_Latn - eng_Latn	deu_Latn - eng_Latn	deu_Latn - fra_Latn
	deu_Latn - fra_Latn	deu_Latn - swe_Latn	deu_Latn - fra_Latn	deu_Latn - nld_Latn	eng_Latn - swe_Latn	ces_Latn - pol_Latn
	dan_Latn - deu_Latn	dan_Latn - deu_Latn	fra_Latn - ind_Latn	ces_Latn - rus_Cyrl	eng_Latn - spa_Latn	deu_Latn - nld_Latn
Unique languages	swe, dan, fra, eng, ukr, rus, zho, yue, deu, spa	fra, eng, swe, dan, rus, ukr, deu, zho, yue, spa	rus, ukr, fra, eng, zho, yue, ind, spa, deu, ces	yue, zho, ukr, rus, swe, dan, fra, eng, spa, deu	zho, yue, rus, ukr, dan, swe, fra, eng, spa, deu	yue, zho, dan, swe, ukr, rus, spa, fra, eng, deu

Table A2: Top correlated language pairs from the layers with peak ANC scores and the unique languages from the top language pairs. Most correlated pairs among LLMs are similar on their HRLs. Despite differing rankings, instruction-tuned LLMs exhibit similar sets of top language pairs with its pre-trained counterparts.

E Observation of Interlingual Alignment Preservation in T-SNE Projections

Through our single-language training experiments in the multilingual mathematical reasoning task, we observe that the visual projections using t-SNE, also support that ILO score effectively captures the same interlingual alignment phenomenon, albeit in a projected lower-dimensional dimensions. In other words, layers with high ILO scores consistently exhibits interlingual overlaps in the t-SNE dimensions that hints at strong interlingual alignment, whereas those with lower scores tend to be more fragmented. This correspondence validates ILO as a robust quantitative measure that reflects the local structure of the multilingual shared embedding space. We attach the complete t-SNE visualization projected from the hidden-states of the models underwent single-language training on English in the **fine-tuning** vs **selective freezing** modes, frozen on their first 8 layers, the token embedding, final layer normalization, and the language modeling head (output projection layers), of Llama-3.1 (8B) and Gemma-2 (9B) respectively in Figures A14 vs A15, and A16 vs A17.

F Ablation Studies

Here we provide comprehensive ablations to all of the hyperparameters in our study and thoroughly analyzes the impact on each of them.

F.1 t-SNE perplexity

We conducted additional t-SNE analysis using perplexity values of 5, 30, and 50, on early, middle, and late layers of Aya Expanse (8B), and visualize them in Figures A18, A19, A20, and A21. Throughout the various perplexity settings, we similarly observe that in the early and late layers, language representations exhibit a minimal overlap, while they cluster according to resource levels and linguistic features. There are different overlaps in the early layer, between Germany and English instead of Japanese and Chinese, when the perplexity is set to 50; additional overlaps between pairs of Bengali, Sinhala, and Czech, Polish in the late layer, with the perplexity set to 5; and no overlap at all in the early layer when the perplexity is set to 30. We also observed similar interlingual overlaps in the intermediate layer that mainly involve high-resource languages with some representations consistently remaining fragmented outside these overlaps, and that low-resource languages overlap due to regional factors. The same set of languages overlaps, with minor differences: the languages of Danish, Swedish, and Ukrainian are added to the overlap with the perplexity set to 5, 30, and 50, and with Yue Chinese missing in the overlaps when the perplexity is set to 50.

These observations substantiate the findings that the interlingual overlapping patterns remain consistent in all cases regardless of the perplexity values used. These additional analyses reinforce the notion that these representational patterns are inherent to the model's learned structure rather than artifacts of a specific t-SNE configuration.

Method	Training languages	Accuracy											Average	
		ben	tha*	swh	tel*	jpn	zho	deu	fra	rus	spa	eng	All	XL
Pre-trained	mixed	13.2%	12.0%	9.2%	16.0%	10.0%	17.6%	16.8%	16.8%	10.8%	15.2%	17.6%	11.2%	-
Fine-tuning	ben	**27.6%**	4.4%	2.0%	4.4%	11.6%	12.8%	6.8%	10.4%	10.0%	14.4%	18.4%	11.2%	9.5%
	tha*	5.6%	**32.4%**	6.0%	2.8%	10.4%	14.4%	14.8%	16.8%	12.0%	20.0%	26.0%	14.7%	12.9%
	swh	5.6%	5.6%	**32.4%**	0.8%	10.4%	9.6%	15.6%	14.8%	10.8%	21.2%	26.4%	13.9%	12.1%
	jpn	2.4%	6.0%	2.8%	2.4%	**26.8%**	19.6%	13.2%	10.8%	14.4%	18.0%	26.0%	12.9%	11.6%
	zho	2.0%	6.4%	1.6%	0.8%	16.8%	**32.0%**	17.6%	10.4%	16.4%	18.0%	28.0%	13.6%	11.8%
	deu	4.4%	9.2%	5.2%	**6.8%**	16.0%	18.4%	**32.8%**	23.6%	23.2%	26.4%	34.4%	18.2%	16.8%
	fra	5.6%	10.8%	6.0%	0.8%	17.6%	18.8%	29.2%	**30.8%**	21.6%	29.6%	31.6%	18.4%	17.2%
	rus	4.8%	4.8%	5.2%	1.2%	13.2%	16.8%	30.0%	24.4%	**32.8%**	29.2%	29.2%	17.4%	15.9%
	spa	7.2%	7.6%	4.8%	4.4%	17.6%	22.0%	26.8%	27.6%	28.4%	**33.2%**	37.6%	**19.7%**	**18.4%**
	eng	8.0%	10.4%	8.0%	6.0%	17.6%	20.8%	28.0%	24.4%	25.2%	29.6%	**39.2%**	**19.7%**	17.8%
Selective Freezing	ben	**36.0%**	13.2%	17.2%	20.0%	22.8%	19.6%	19.6%	22.0%	21.2%	18.0%	26.8%	21.5%	20.0%
	tha*	14.4%	**34.4%**	14.0%	13.6%	16.8%	21.6%	20.0%	22.8%	21.2%	24.8%	27.2%	21.0%	19.6%
	swh	13.2%	14.4%	**30.4%**	11.2%	15.2%	20.4%	26.8%	25.2%	20.8%	29.6%	29.6%	21.5%	20.6%
	jpn	12.8%	14.8%	19.2%	13.2%	**27.6%**	26.8%	22.0%	21.6%	23.6%	21.6%	26.4%	20.9%	20.2%
	zho	12.8%	19.2%	15.6%	13.6%	22.0%	**34.8%**	26.4%	27.2%	22.4%	24.8%	31.2%	22.7%	21.5%
	deu	11.2%	17.6%	18.8%	14.0%	20.0%	21.2%	33.6%	26.0%	26.8%	28.0%	35.2%	22.9%	21.9%
	fra	20.4%	17.6%	22.4%	20.0%	23.6%	24.0%	30.4%	35.6%	28.4%	33.2%	32.8%	26.2%	25.3%
	rus	15.2%	17.6%	24.0%	17.2%	18.4%	18.4%	28.8%	26.0%	**36.4%**	27.6%	32.4%	23.8%	22.6%
	spa	18.4%	21.2%	26.4%	18.8%	22.0%	26.4%	**36.4%**	31.6%	29.2%	35.6%	38.8%	27.7%	26.9%
	eng	22.4%	25.6%	26.8%	**22.4%**	24.8%	26.0%	34.4%	**36.0%**	34.0%	**39.2%**	**41.6%**	**30.3%**	**29.2%**

Table A3: Cross-lingual transfer performance on MGSM for Gemma-2 (9B) w/ and w/o selective freezing. "XL" denotes average on languages that were not fine-tuned. Diagonal entries in blue highlights correspond to source language performances. Red highlights indicate decrease from pre-trained baseline. **Bold** and underline respectively denote the best within group and within column. The (*) marks languages classified as low-resource in Flores-200.

F.2 k-NN parameters of the ILO score

We further conducted ablation studies over different settings of k and τ—specifically, [(5,3), (10,5), (20,10)]—using both cosine and Euclidean distances. We report the results in Table A22 and A23. Our results indicate that a lower k ($k = 5$, $\tau = 3$) leads to a modest increase in the overall aggregated ILO across all layers by about 0.03–0.05, whereas a higher k ($k = 20$, $\tau = 10$) results in a reduction of roughly 0.1–0.15 relative to our main illustration in Figure 5. Nonetheless, we find that all the trends remain consistent with our findings. When ablating a different distance metric, i.e, cosine distance, we find that the influence of varying k values is slightly less pronounced, with the aggregated ILO scores remaining within a similar range.

In summary, despite the different selection of the k-NN parameters and distance metric, observations using ILO score consistently highlight similar trend on the decrease of alignment degree in the same layers, and that the model trained with the selective-freezing mechanism sustains their prior semantic alignment levels in all layers.

F.3 Layer selection for selective freezing

We perform experiments on selective freezing of the first 4, 8, 12, and 16 layers of Llama-3.1 (8B). Our motivation stems from prior works that have demonstrated that multilingual language models tend to align their representations in the early layers (Muller et al., 2021; Zhao et al., 2024), which guided our decision to focus on these layers. We denote the aggregated results in Table A4, the complete results in Table A5, and visualize the aggregated ILO scores in Figures A24. In general, fine-tuning with freezing the early layers enhances the cross-lingual generalization. Notably, the best overall performance was achieved when freezing the first 12 layers. Throughout the experiments, analysis of interlingual alignment using ILO reveal that freezing the first 4, 8, and 12 layers maintains and improves the semantic alignment across layers. In contrast, while freezing the first 16 layers preserves alignment in the frozen layers, the subsequent layers exhibit lower alignments compared to the fine-tuned models.

Furthermore, across all settings, we observed improved transfer on languages within families and regions, with negligible degradation—and sometimes even improvements—in low-resource, cross-family, and cross-regional scenarios. When comparing the trade-offs between freezing the first 8 layers versus the first 12 layers, we found that the performance gain in the source language is

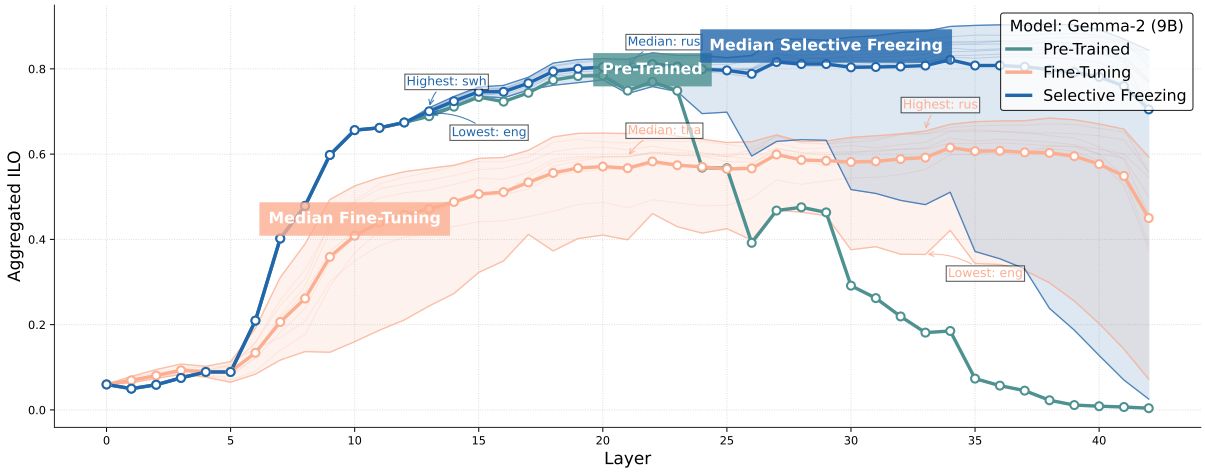


Figure A1: Layer-wise $\overline{\text{ILO}}_{\mathcal{L}}$ scores for Gemma-2 (9B) in **pre-trained**, **fine-tuning**, and **selective freezing** modes. Notable decrease in alignment from single-language training is seen in the early layers on **fine-tuning**, whereas the **selective freezing** mechanism allows the model to sustain its **pre-trained** semantic alignment across layers.

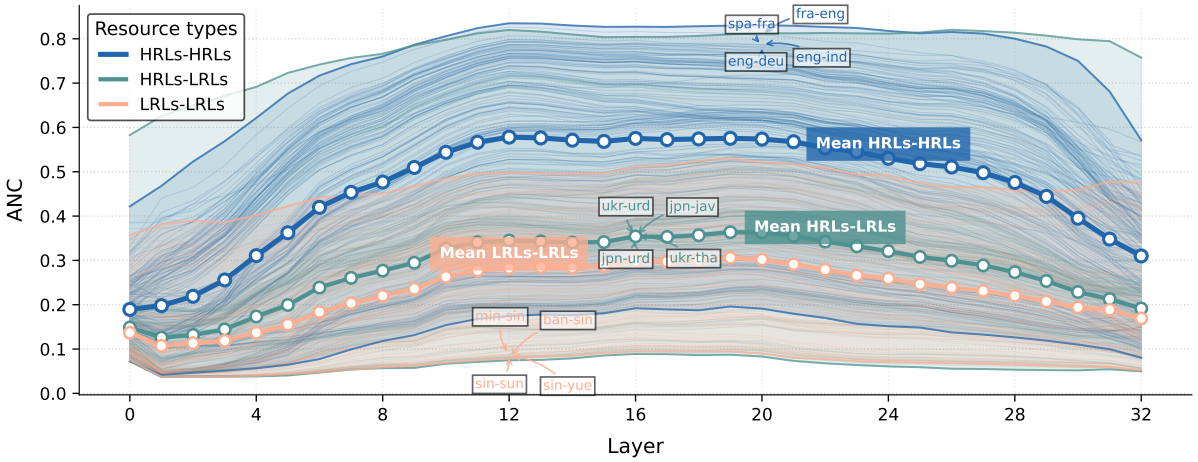
Method	**Frozen Layers**	**Average**	
		All	**XL**
Fine-tuning	-	17.7%	16.0%
Selective Freezing	First 4	21.6%	20.2%
	First 8	22.4%	21.2%
	First 12	**23.1%**	**22.1%**
	First 16	19.0%	18.0%

Table A4: Aggregated results on the ablation study on the cross-lingual transfer performance on MGSM for Llama-3.1 (8B) fine-tuned with the selective freezing strategy varied on the frozen layers. Freezing the first 4, 8, 12, and 16 layers enhanced the cross-lingual generalization, with the best performance achieved when freezing the first 12 layers.

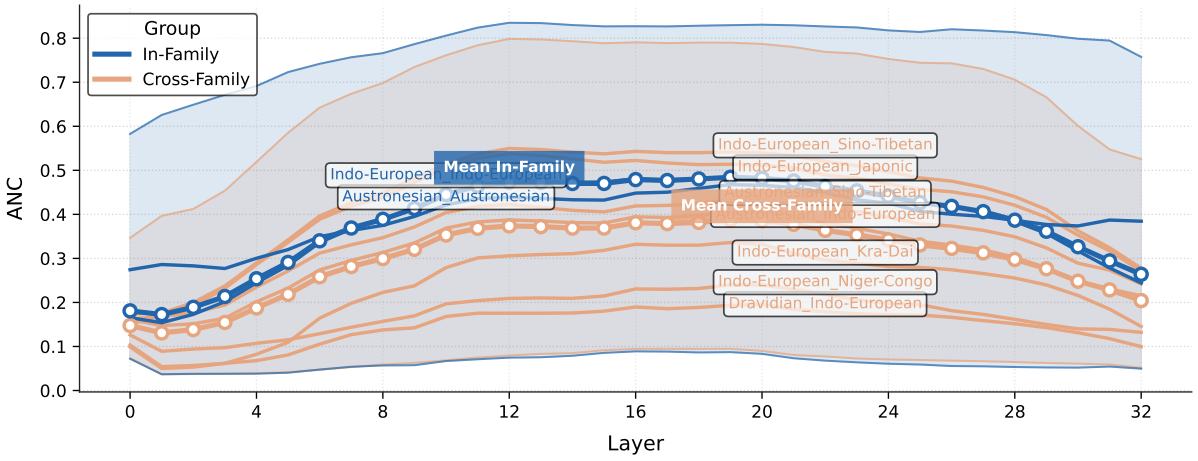
mixed. In the latter setting, the task performances in languages such as English, Russian, French, German, and Bengali improved, while in Spanish, Chinese, Japanese, Swahili, and Thai, they instead decreased. Moreover, the multilingual performance from fine-tuning with English mostly dropped, except for certain gains in English, Russian, Thai, and Bengali. Lastly, the aggregate multilingual performance when freezing the first 16 layers is closer to that of fine-tuning, showcasing the impact of lower interlingual alignment previously indicated from the observation on the analysis using ILO.

Method	Training languages	Accuracy											Average	
		ben	tha*	swh	tel*	jpn	zho	deu	fra	rus	spa	eng	All	XL
Pre-trained	mixed	11.6%	12.0%	7.2%	0.0%	10.4%	8.8%	16.0%	12.4%	14.0%	11.6%	17.6%	10.3%	-
Fine-tuning	ben	**23.2%**	4.8%	1.2%	3.2%	10.0%	9.6%	10.8%	13.6%	11.6%	14.8%	12.8%	10.5%	9.2%
	tha*	1.6%	**32.8%**	4.4%	1.6%	14.4%	14.8%	17.2%	19.2%	18.0%	20.4%	25.6%	15.5%	13.7%
	swh	3.2%	6.4%	**30.8%**	2.8%	11.2%	12.4%	20.4%	19.6%	14.8%	22.4%	26.8%	15.5%	14.0%
	jpn	3.6%	7.2%	2.8%	1.2%	**32.8%**	21.6%	19.6%	18.0%	18.4%	22.4%	28.8%	16.0%	14.4%
	zho	0.8%	7.2%	2.4%	1.6%	22.0%	**34.8%**	19.6%	19.6%	21.6%	21.2%	27.6%	16.2%	14.4%
	deu	8.0%	16.4%	8.0%	**4.0%**	19.2%	19.6%	**37.6%**	**34.4%**	23.6%	28.8%	36.4%	21.5%	19.8%
	fra	4.8%	11.6%	4.0%	3.2%	16.0%	16.8%	31.6%	**34.4%**	25.6%	34.4%	35.6%	19.8%	18.4%
	rus	4.0%	14.0%	4.0%	1.2%	17.2%	16.4%	29.6%	28.4%	**34.0%**	30.0%	26.4%	18.7%	17.1%
	spa	4.8%	16.0%	2.8%	2.4%	14.4%	19.6%	28.4%	30.8%	31.2%	**38.4%**	38.4%	20.7%	18.9%
	eng	6.4%	14.4%	6.0%	2.4%	18.8%	24.4%	37.2%	27.2%	33.6%	33.2%	**43.2%**	**22.4%**	**20.4%**
Selective Freezing First 4 Layers	ben	**22.8%**	8.8%	8.0%	12.4%	14.8%	10.0%	12.0%	12.4%	14.4%	16.4%	14.8%	13.3%	12.4%
	tha*	10.4%	**31.2%**	5.2%	9.2%	17.2%	20.0%	19.6%	18.4%	15.2%	19.6%	28.8%	17.7%	16.4%
	swh	9.6%	15.2%	**38.4%**	11.2%	11.6%	17.6%	26.0%	23.2%	16.4%	28.0%	26.4%	20.3%	18.5%
	jpn	14.4%	12.0%	10.4%	11.2%	**36.4%**	24.8%	23.2%	19.2%	24.8%	19.6%	25.2%	20.1%	18.5%
	zho	11.6%	15.6%	10.8%	6.8%	20.0%	**36.0%**	27.6%	26.0%	19.6%	29.2%	29.2%	21.1%	19.6%
	deu	14.8%	20.8%	10.4%	10.4%	14.8%	19.6%	**38.8%**	32.0%	27.2%	31.2%	38.4%	23.5%	22.0%
	fra	14.8%	18.8%	8.8%	10.0%	20.8%	23.6%	34.4%	**38.0%**	31.6%	35.6%	37.6%	24.9%	23.6%
	rus	15.6%	16.4%	10.0%	10.4%	21.2%	20.8%	27.6%	26.8%	38.0%	26.0%	37.6%	22.8%	21.2%
	spa	15.6%	17.2%	11.2%	8.0%	20.4%	21.6%	31.6%	32.8%	34.4%	**38.0%**	35.6%	24.2%	22.8%
	eng	17.2%	25.6%	13.2%	**13.2%**	23.6%	28.0%	36.0%	34.8%	**38.8%**	36.4%	**41.2%**	**28.0%**	**26.7%**
Selective Freezing First 8 Layers	ben	**23.2%**	9.2%	8.8%	10.0%	17.6%	11.6%	18.0%	16.4%	17.6%	18.4%	20.8%	15.6%	14.8%
	tha*	14.0%	**35.2%**	12.4%	12.4%	16.4%	20.8%	24.8%	20.8%	16.8%	18.0%	28.0%	20.0%	18.4%
	swh	8.4%	13.6%	**30.0%**	8.4%	15.2%	12.8%	20.8%	19.2%	16.8%	24.8%	29.2%	18.1%	16.9%
	jpn	15.6%	15.2%	12.0%	14.0%	**30.0%**	27.2%	24.8%	22.8%	23.2%	24.0%	28.0%	21.5%	20.7%
	zho	15.6%	21.2%	10.4%	10.4%	22.0%	**40.8%**	23.6%	20.4%	21.6%	25.2%	34.8%	22.4%	20.5%
	deu	18.0%	18.4%	8.4%	16.0%	22.4%	24.0%	34.0%	31.2%	27.6%	32.0%	38.4%	24.6%	23.6%
	fra	**23.2%**	19.2%	13.2%	14.0%	18.8%	20.0%	30.4%	**35.2%**	30.8%	33.2%	37.6%	25.1%	24.0%
	rus	17.2%	18.4%	10.8%	14.4%	15.2%	18.0%	29.6%	24.4%	**38.0%**	29.6%	36.8%	22.9%	21.4%
	spa	17.2%	18.4%	11.6%	14.0%	20.4%	22.8%	31.6%	31.6%	28.8%	38.0%	36.4%	24.6%	23.3%
	eng	18.8%	23.2%	19.6%	**17.6%**	26.4%	29.6%	**36.8%**	32.4%	36.4%	**40.0%**	**42.0%**	**29.3%**	**28.1%**
Selective Freezing First 12 Layers	ben	**26.4%**	12.8%	11.6%	14.4%	13.6%	14.8%	19.6%	20.0%	20.0%	17.6%	17.2%	17.1%	16.2%
	tha*	14.8%	**34.0%**	12.0%	12.4%	15.6%	21.6%	25.2%	22.0%	20.4%	24.4%	32.4%	21.3%	20.1%
	swh	9.2%	16.4%	**22.8%**	5.6%	14.0%	12.4%	18.4%	23.6%	19.2%	20.4%	27.6%	17.2%	16.7%
	jpn	16.0%	17.6%	12.0%	11.2%	**27.2%**	28.8%	24.4%	23.2%	24.0%	24.4%	29.6%	21.7%	21.1%
	zho	17.2%	17.2%	12.4%	12.0%	22.4%	**34.8%**	29.6%	22.4%	27.6%	23.6%	37.2%	23.3%	22.2%
	deu	12.8%	22.8%	14.4%	17.6%	20.0%	25.6%	**36.0%**	29.6%	27.6%	32.8%	39.2%	25.3%	24.2%
	fra	14.8%	24.8%	18.4%	12.0%	21.2%	21.2%	33.6%	**37.2%**	32.0%	**36.8%**	36.8%	26.3%	25.2%
	rus	20.4%	19.6%	11.6%	**18.8%**	22.0%	19.6%	28.8%	25.2%	38.4%	28.8%	32.0%	24.1%	22.7%
	spa	20.0%	24.0%	17.6%	16.8%	18.0%	27.2%	33.6%	33.6%	29.6%	34.0%	36.4%	26.4%	25.7%
	eng	20.4%	24.0%	18.0%	16.4%	20.4%	26.4%	35.2%	30.0%	**43.6%**	32.4%	**46.8%**	**28.5%**	**26.7%**
Selective Freezing First 16 Layers	ben	**24.0%**	13.6%	6.4%	10.4%	11.2%	7.6%	16.8%	16.0%	15.2%	13.6%	16.0%	13.7%	12.7%
	tha*	11.6%	**27.2%**	9.6%	10.4%	12.4%	15.6%	19.6%	14.4%	21.2%	19.6%	27.6%	17.2%	16.2%
	swh	10.8%	10.8%	**20.4%**	8.0%	11.6%	10.4%	18.0%	20.4%	14.8%	19.6%	21.2%	15.1%	14.6%
	jpn	14.8%	13.6%	9.6%	6.0%	**26.4%**	22.4%	23.2%	17.2%	14.8%	22.0%	26.8%	17.9%	17.0%
	zho	12.8%	15.2%	6.0%	8.0%	15.6%	**27.2%**	23.2%	16.0%	24.0%	21.6%	31.6%	18.3%	17.4%
	deu	10.4%	19.6%	9.2%	9.6%	15.6%	20.4%	**34.0%**	23.6%	24.4%	25.2%	34.8%	20.6%	19.3%
	fra	18.4%	14.8%	12.0%	12.8%	14.4%	20.4%	25.6%	**35.6%**	27.6%	30.4%	32.4%	22.2%	20.9%
	rus	12.0%	18.0%	10.0%	12.4%	13.2%	20.0%	26.4%	22.8%	27.6%	23.6%	29.2%	19.6%	18.8%
	spa	11.2%	22.0%	14.0%	**14.8%**	12.8%	20.4%	25.6%	29.2%	30.0%	**30.8%**	32.0%	22.1%	21.2%
	eng	16.0%	16.8%	12.4%	10.4%	17.6%	25.6%	31.2%	30.0%	**34.0%**	26.0%	**40.4%**	**23.7%**	**22.0%**

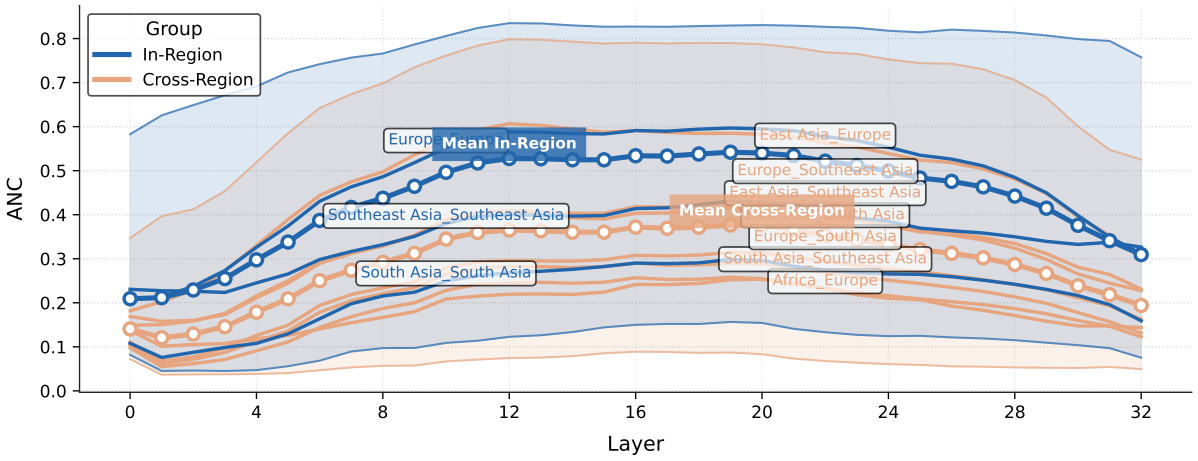
Table A5: Ablation study on the cross-lingual transfer performance on MGSM for Llama-3.1 (8B) fine-tuned with the selective freezing strategy varied on the frozen layers. "XL" denotes average on languages that were not fine-tuned. Diagonal entries in blue highlights correspond to source language performances. Red highlights indicate decrease from pre-trained baseline. **Bold** and underline respectively denote the best within group and within column. The (*) marks languages classified as low-resource in Flores-200.



(a) Highlights on pairs w.r.t their resource levels

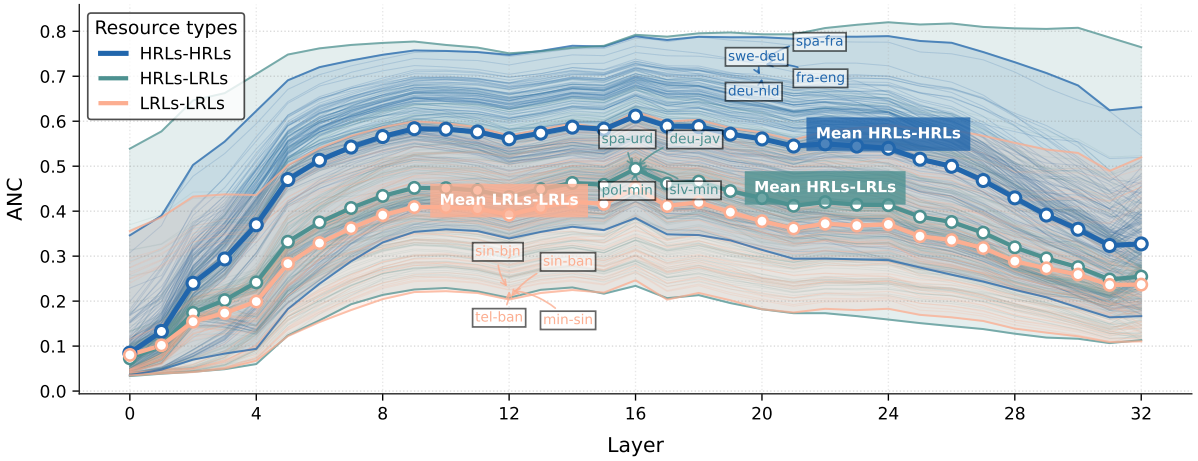


(b) Highlights on pairs w.r.t their linguistic region

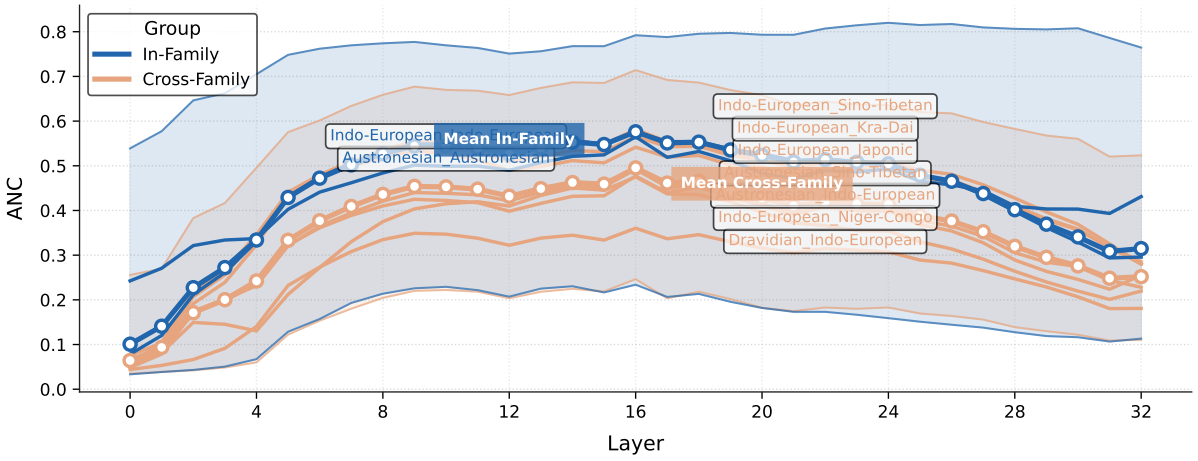


(c) Highlights on pairs w.r.t their linguistic family

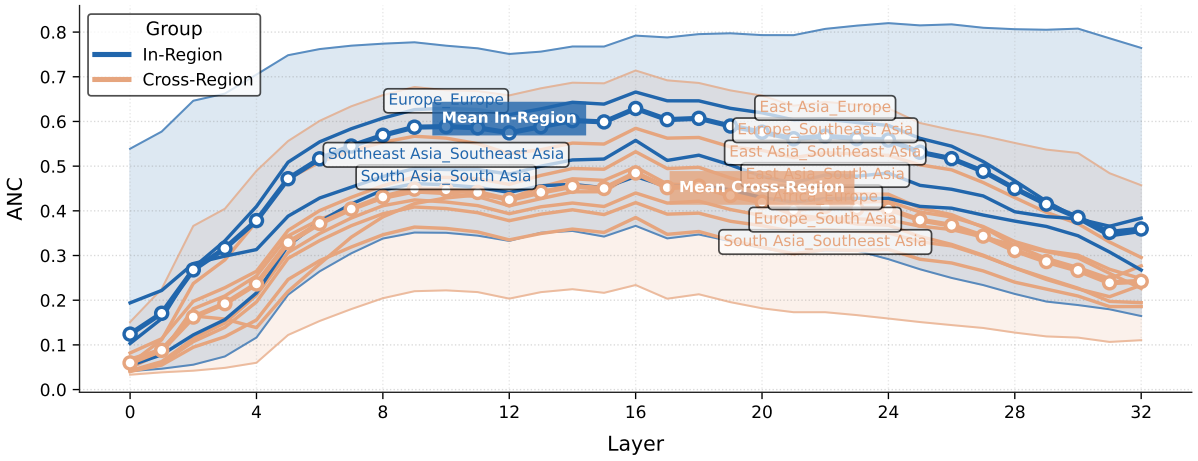
Figure A2: Comparisons of per-layer ANC scores on Aya Expanse (8B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Highlights on pairs w.r.t their resource levels

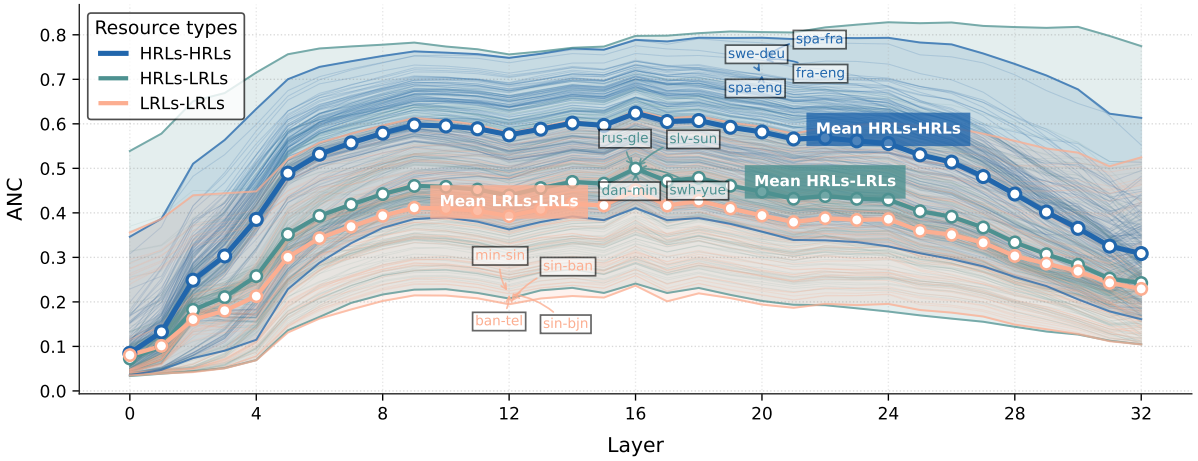


(b) Highlights on pairs w.r.t their linguistic region

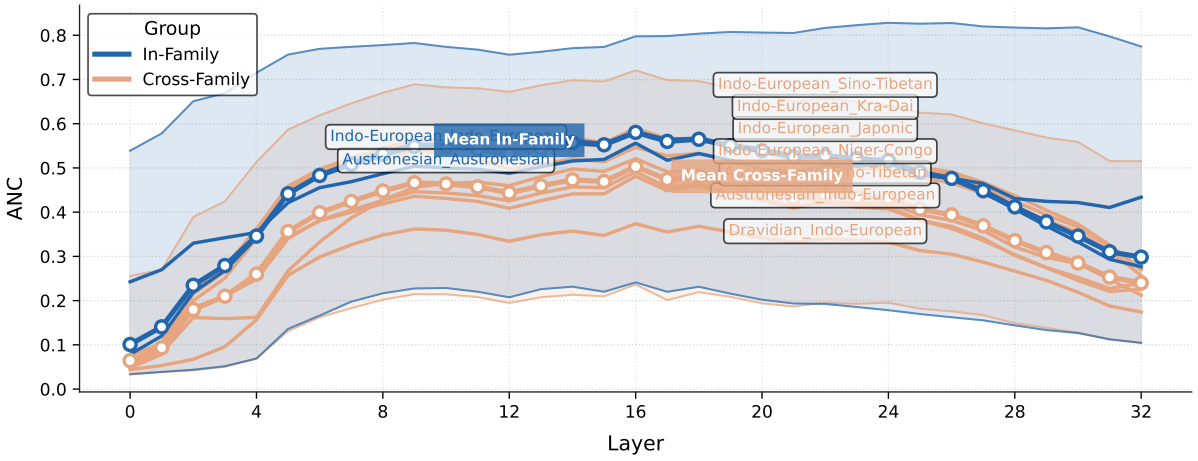


(c) Highlights on pairs w.r.t their linguistic family

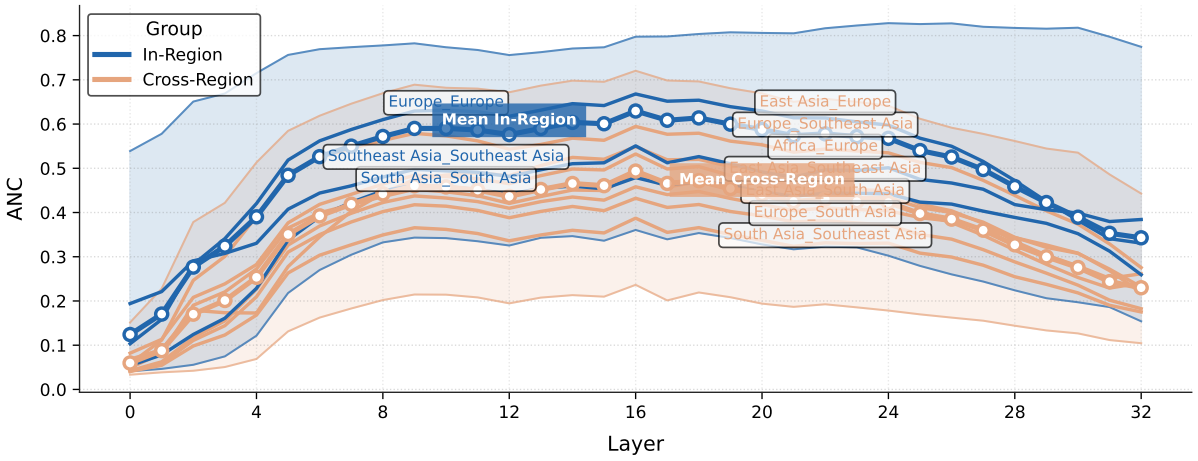
Figure A3: Comparisons of per-layer ANC scores on Llama-3.1 (8B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Highlights on pairs w.r.t their resource levels

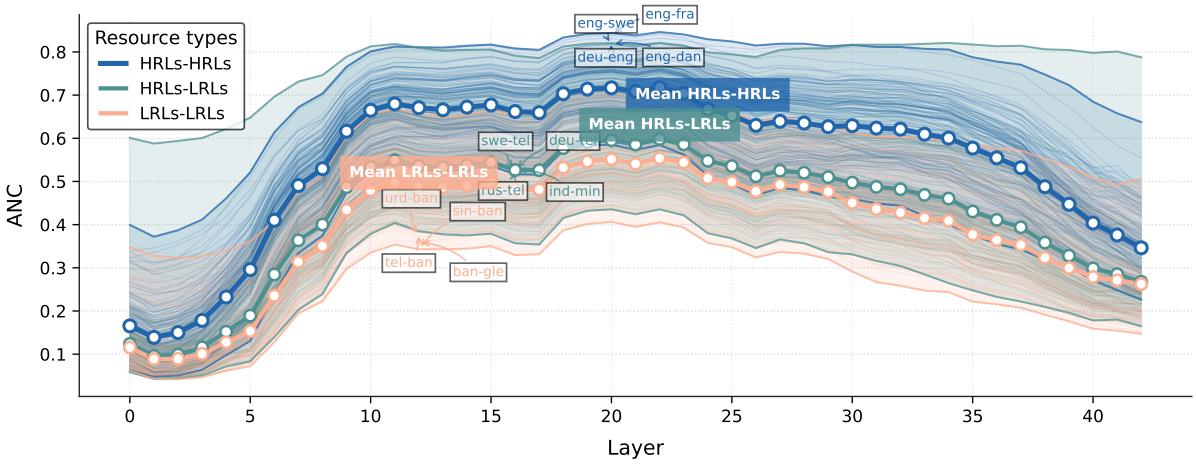


(b) Highlights on pairs w.r.t their linguistic region

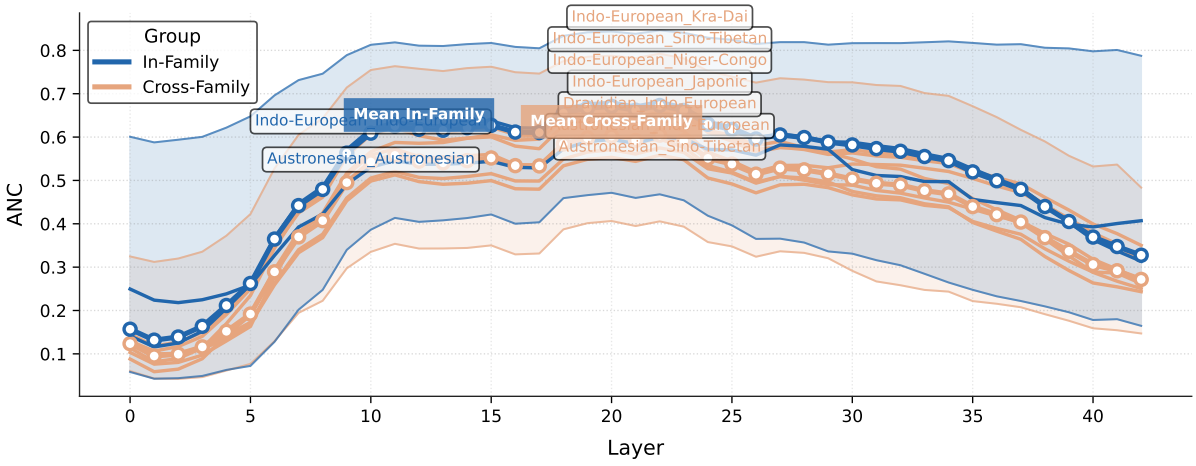


(c) Highlights on pairs w.r.t their linguistic family

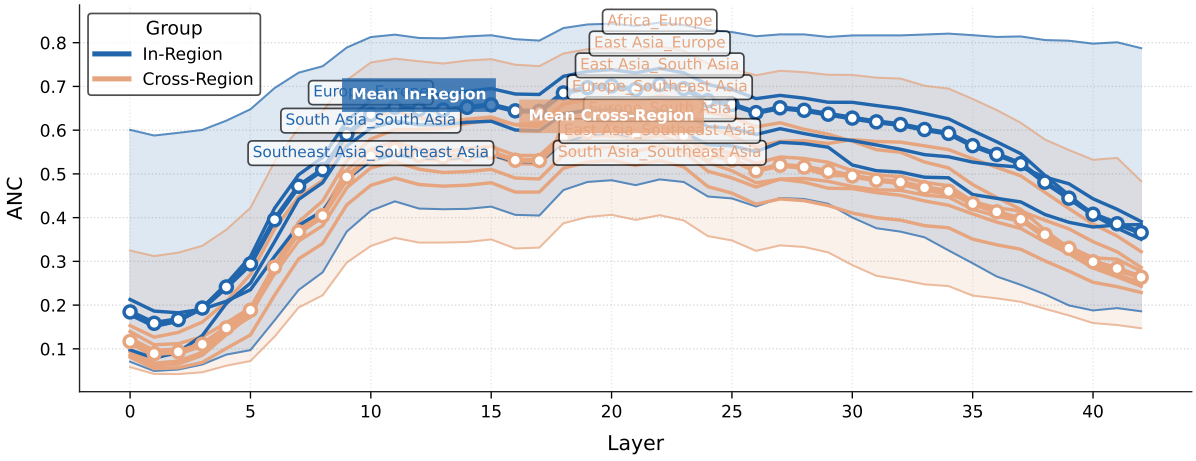
Figure A4: Comparisons of per-layer ANC scores on Llama-3.1-Instruct (8B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Highlights on pairs w.r.t their resource levels

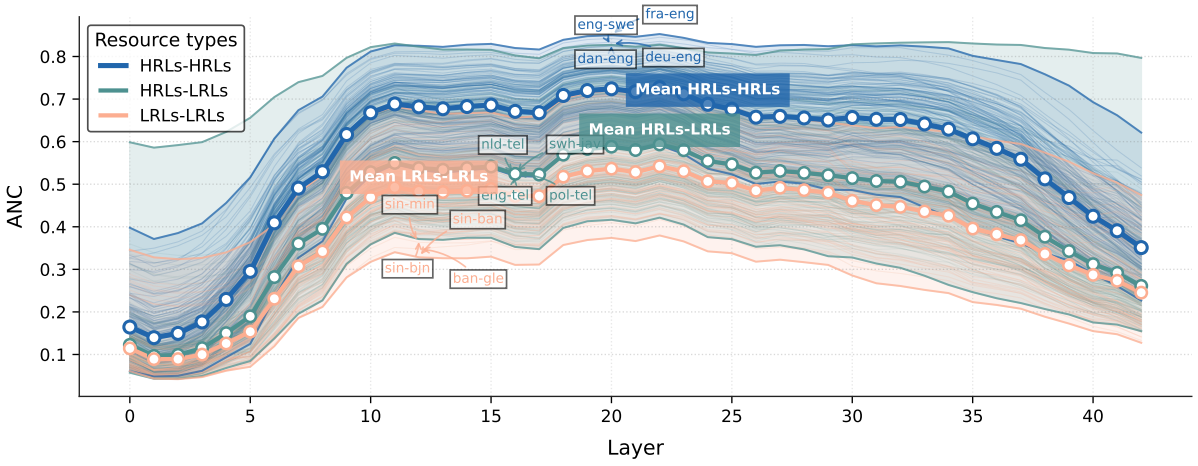


(b) Highlights on pairs w.r.t their linguistic region

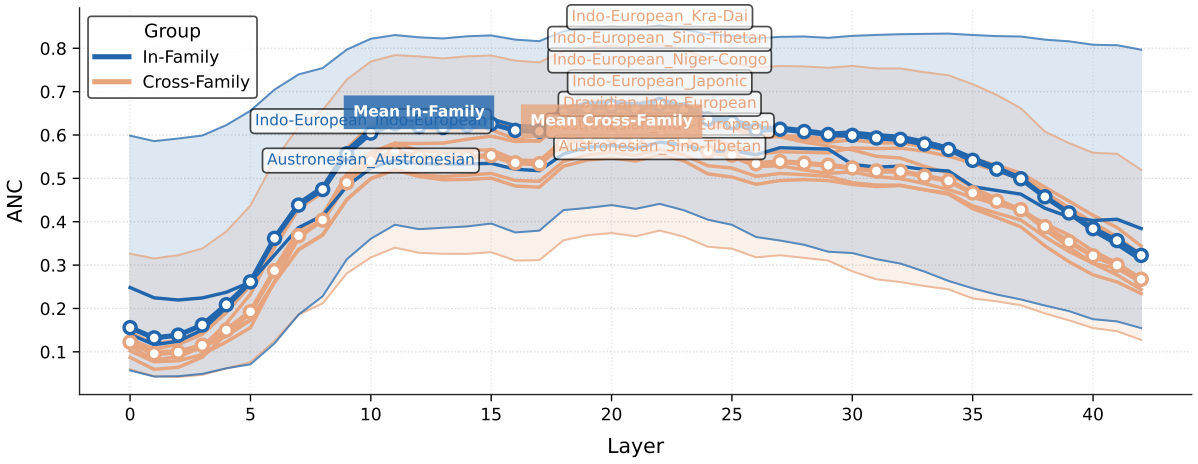


(c) Highlights on pairs w.r.t their linguistic family

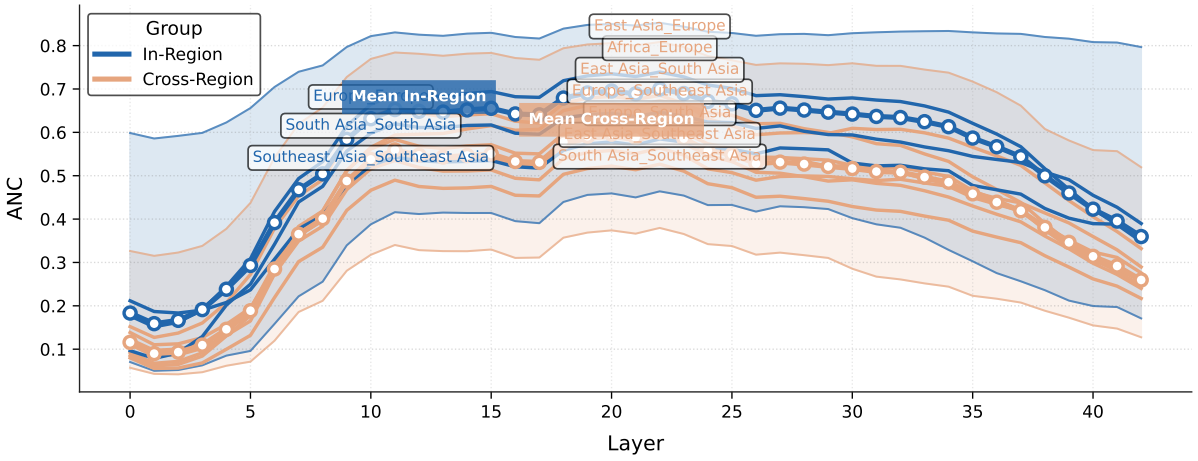
Figure A5: Comparisons of per-layer ANC scores on Gemma-2 (9B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Highlights on pairs w.r.t their resource levels

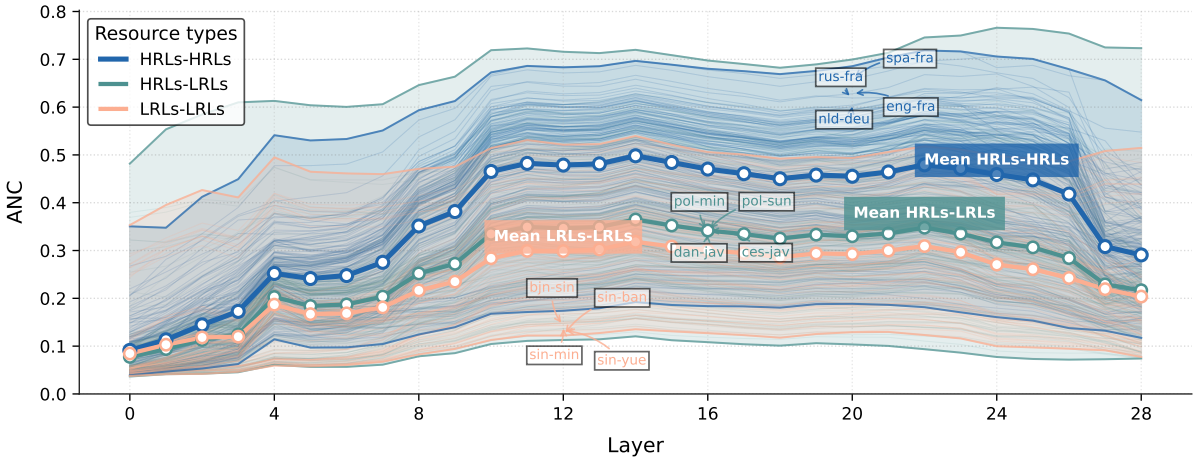


(b) Highlights on pairs w.r.t their linguistic region

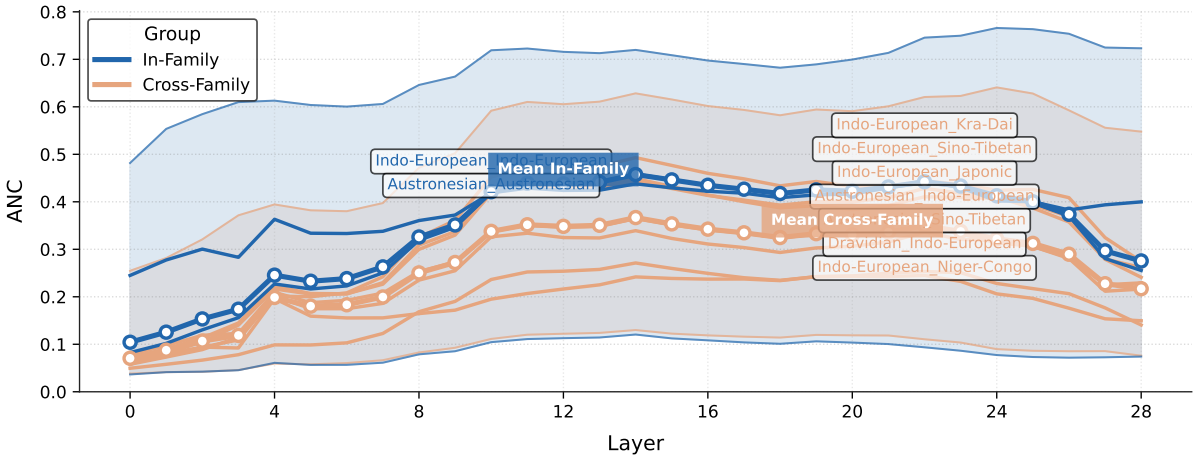


(c) Highlights on pairs w.r.t their linguistic family

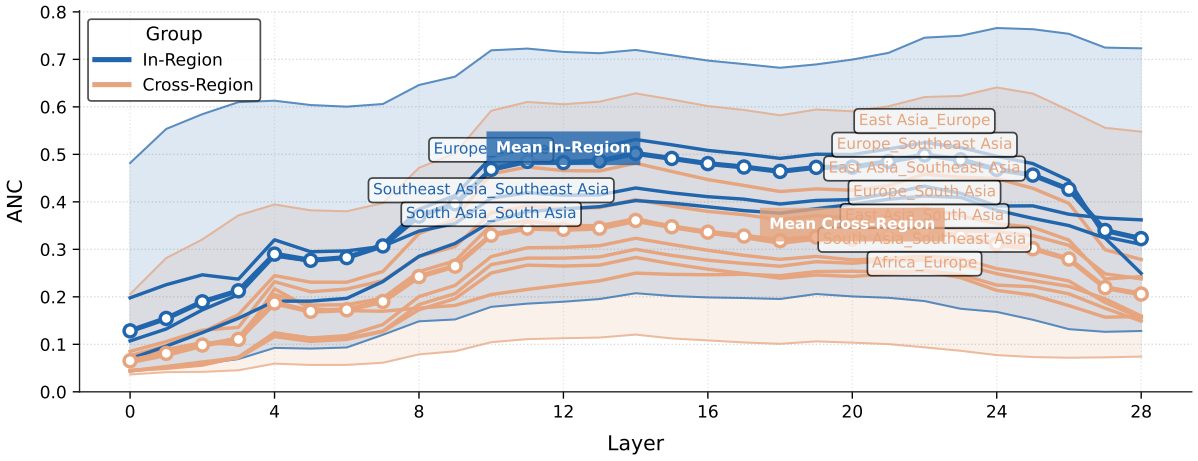
Figure A6: Comparisons of per-layer ANC scores on Gemma-2-Instruct (9B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Highlights on pairs w.r.t their resource levels

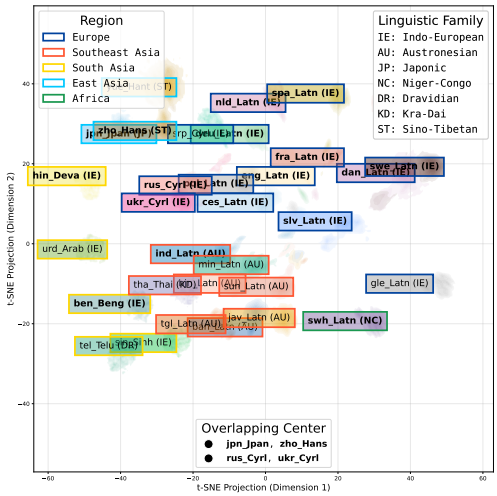


(b) Highlights on pairs w.r.t their linguistic region

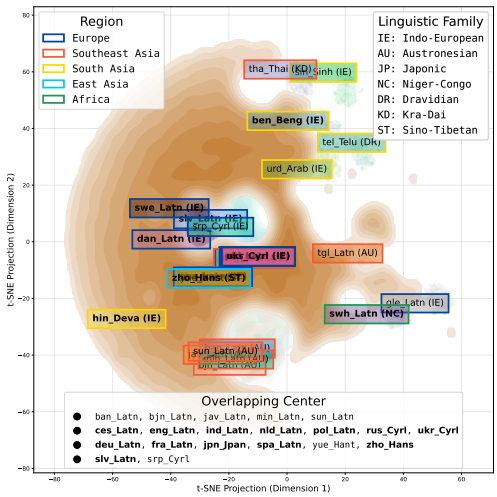


(c) Highlights on pairs w.r.t their linguistic family

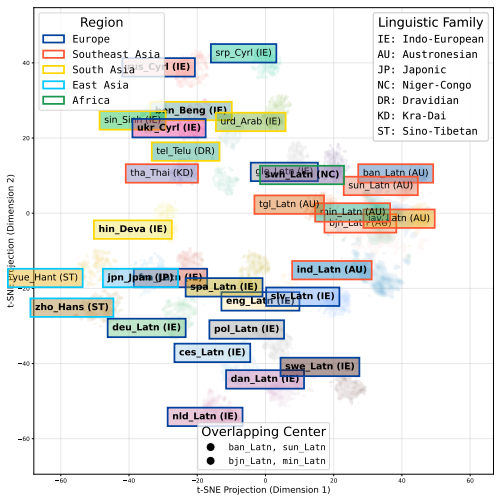
Figure A7: Comparisons of per-layer ANC scores on Qwen-2.5 (7B) with highlights on pairs w.r.t their resource levels, linguistic region and family. Consistently stronger alignments are observed between HRLs pairs and within-group mean correlations.



(a) Early (layer 0)

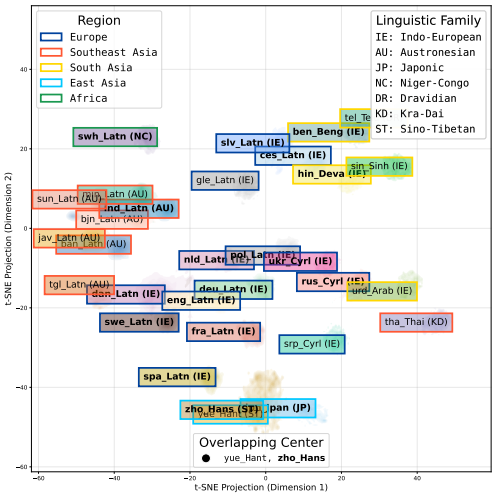


(b) Intermediate (layer 16)

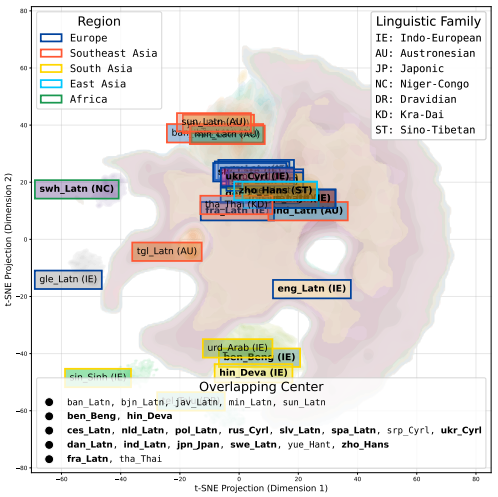


(c) Late (layer 32)

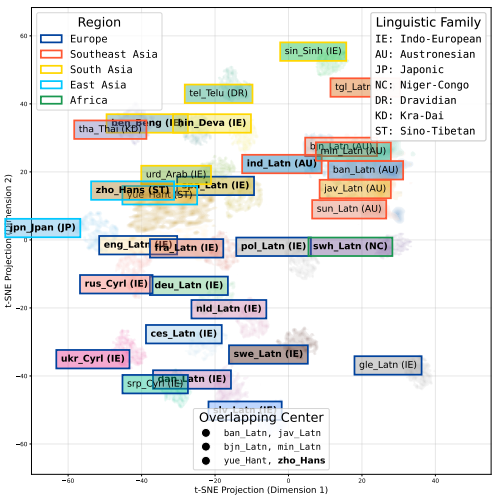
Figure A8: Hidden-state embeddings of Aya Expanse (8B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in the intermediate layer representations. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, with minimal overlap.



(a) Early (layer 0)

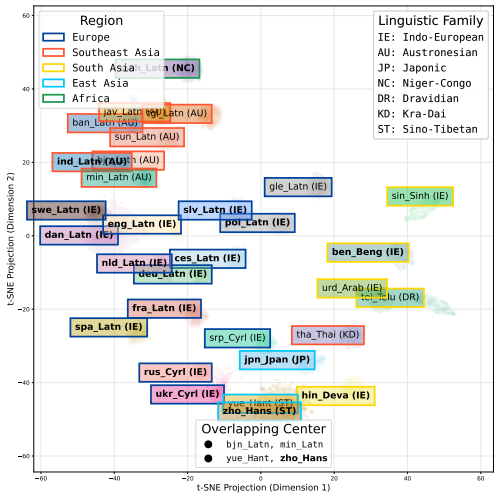


(b) Intermediate (layer 14)

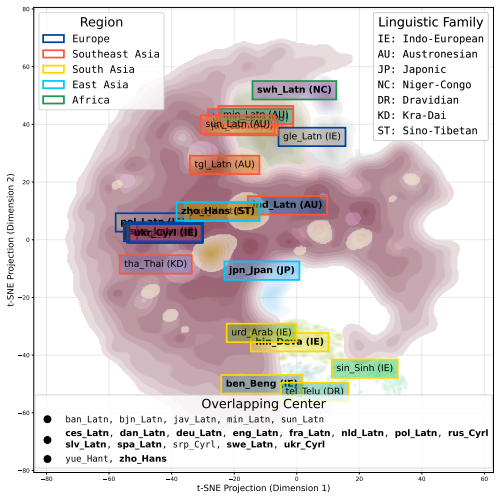


(c) Late (layer 28)

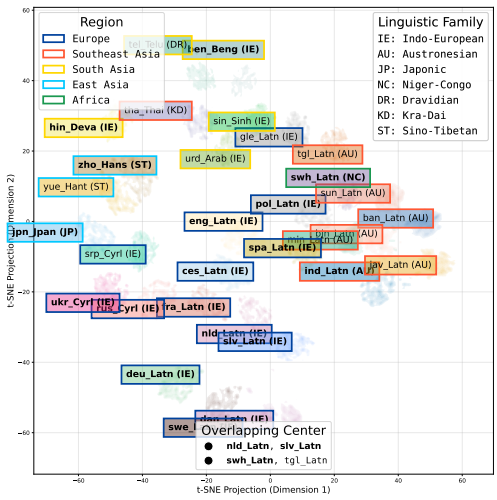
Figure A9: Hidden-state embeddings of Qwen-2.5 (7B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in the intermediate layer representations. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, with minimal overlap.



(a) Early (layer 0)

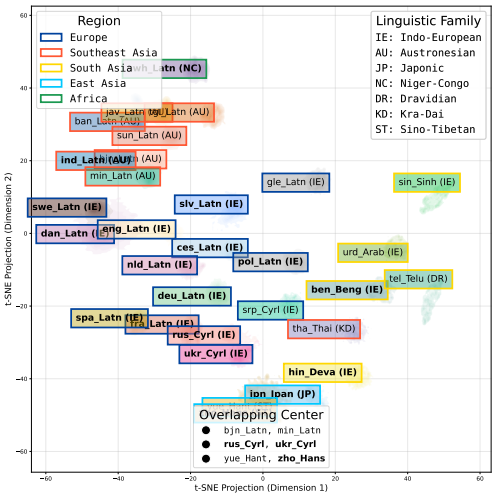


(b) Intermediate (layer 16)

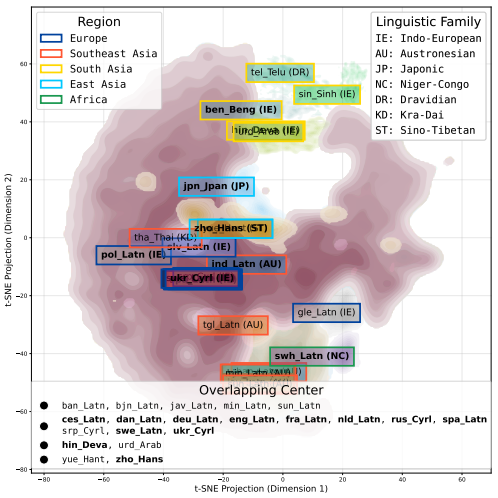


(c) Late (layer 32)

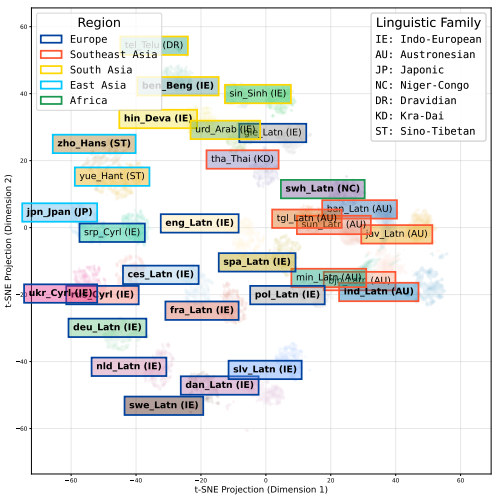
Figure A10: Hidden-state embeddings of Llama-3.1 (8B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in the intermediate layer representations. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, with minimal overlap.



(a) Early (layer 0)

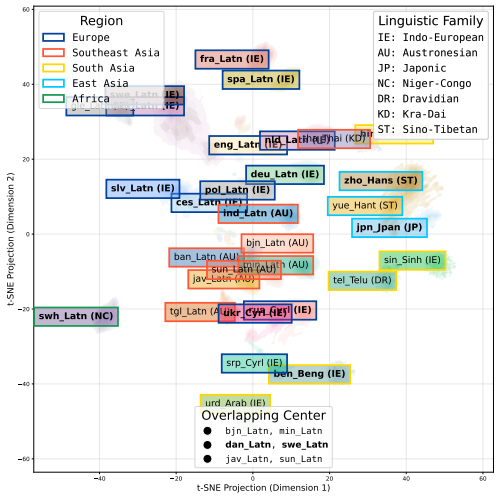


(b) Intermediate (layer 16)

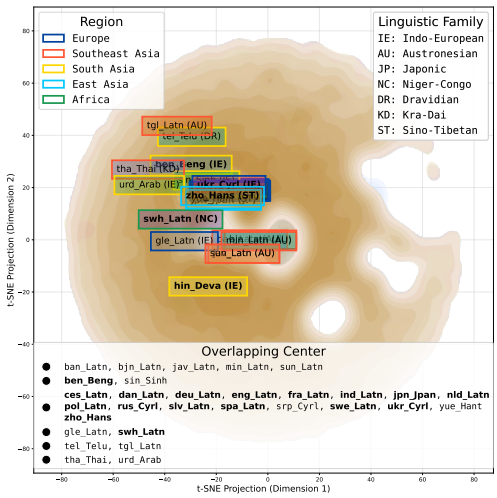


(c) Late (layer 32)

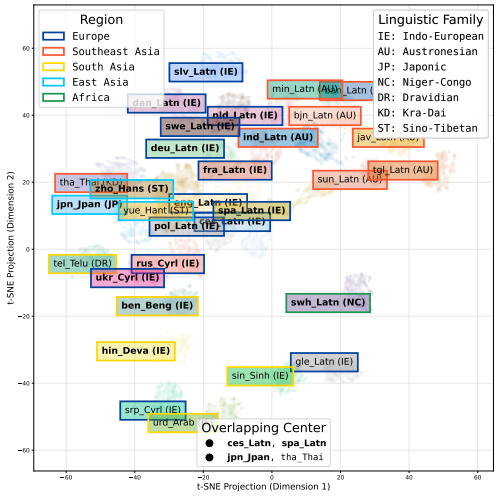
Figure A11: Hidden-state embeddings of Llama-3.1-Instruct (8B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in the intermediate layer representations. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, with minimal overlap.



(a) Early (layer 0)

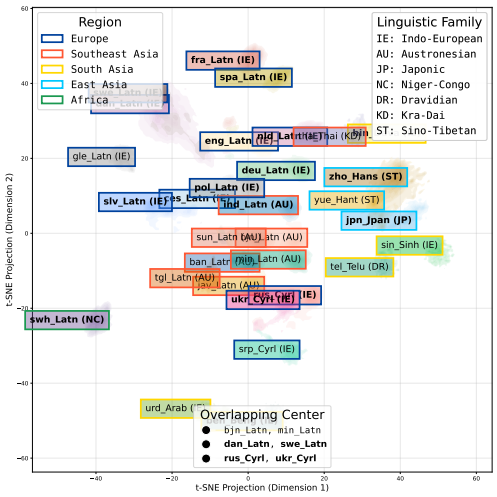


(b) Intermediate (layer 21)

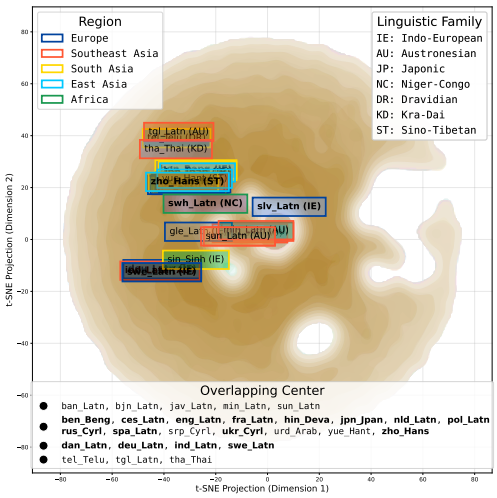


(c) Late (layer 42)

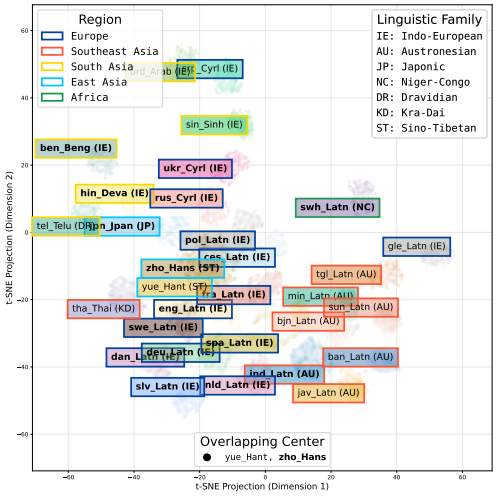
Figure A12: Hidden-state embeddings of Gemma-2 (9B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in the intermediate layer representations. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, with minimal overlap.



(a) Early (layer 0)

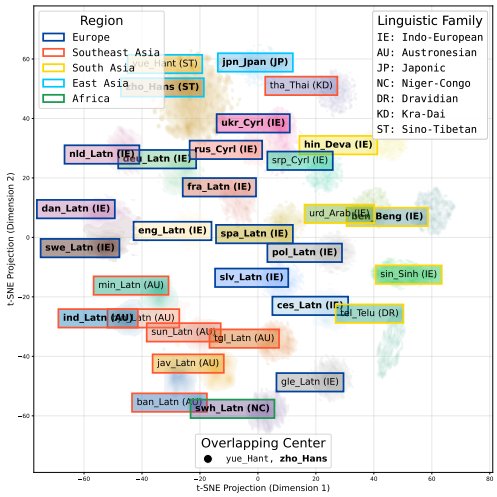


(b) Intermediate (layer 21)

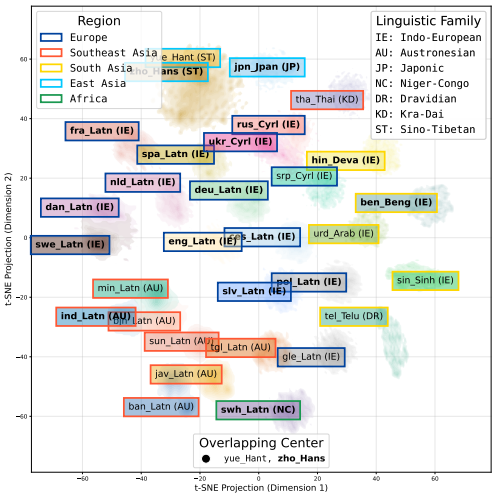


(c) Late (layer 42)

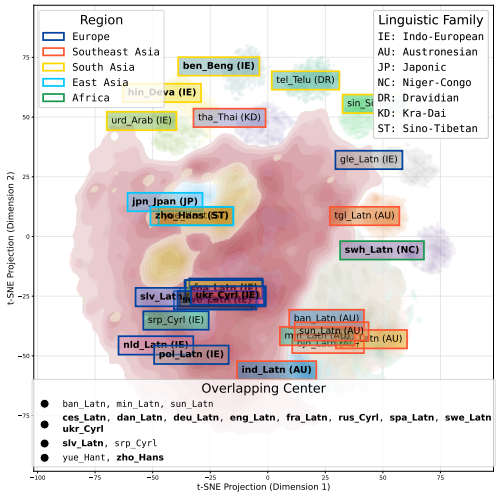
Figure A13: Hidden-state embeddings of Gemma-2-Instruct (9B) projected in t-SNE dimensions, with HRLs in **bold**. Interlingual overlaps transcending familial and regional boundaries are observed in the intermediate layer representations. In the early and late layers, language representations cluster w.r.t resource levels and linguistic features, with minimal overlap.



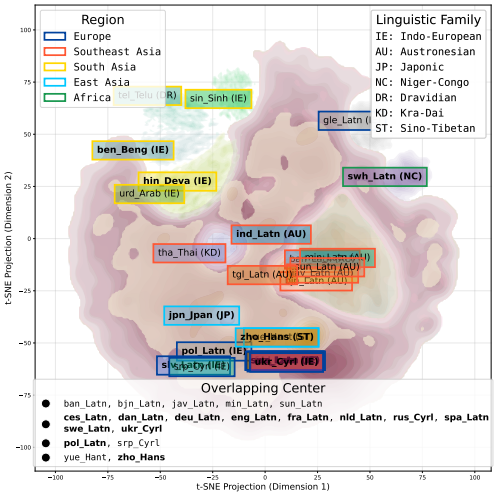
(a) Early (layer 0)



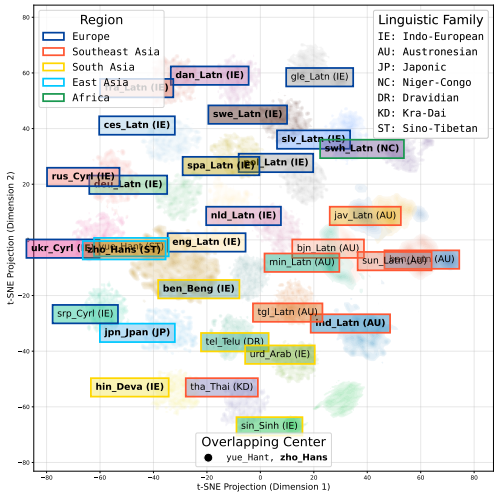
(a) Early (layer 0)



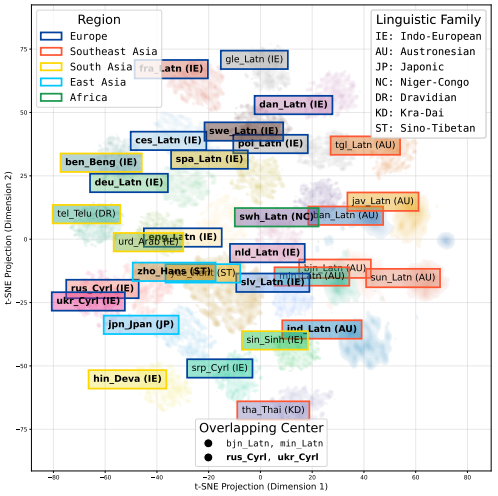
(b) Intermediate (layer 16)



(b) Intermediate (layer 16)



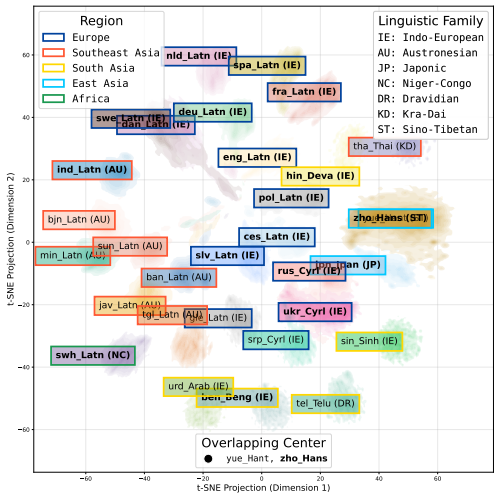
(c) Late (layer 32)



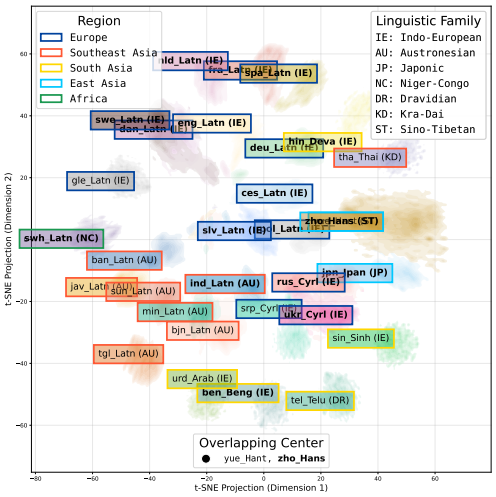
(c) Late (layer 32)

Figure A14: Hidden-state embeddings of Llama-31 (8B) **fine-tuned** on single-language dataset on English, projected in t-SNE dimensions, with HRLs in **bold**. The decline in interlingual semantic alignment is evident from the reduced interlingual overlaps in the projected embeddings within the model's intermediate layer, compared to the observations in Figure A10.

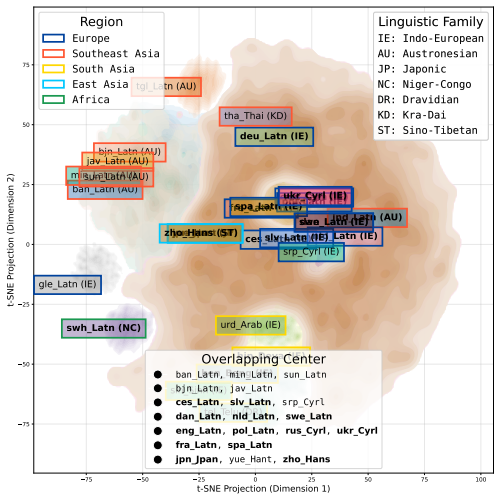
Figure A15: Hidden-state embeddings of Llama-31 (8B) fine-tuned on single-language dataset on English, with **selective freezing** strategy, projected in t-SNE dimensions, with HRLs in **bold**. This approach preserved interlingual alignment, as indicated by high ILO scores that correlate with observed preservation of interlingual overlaps.



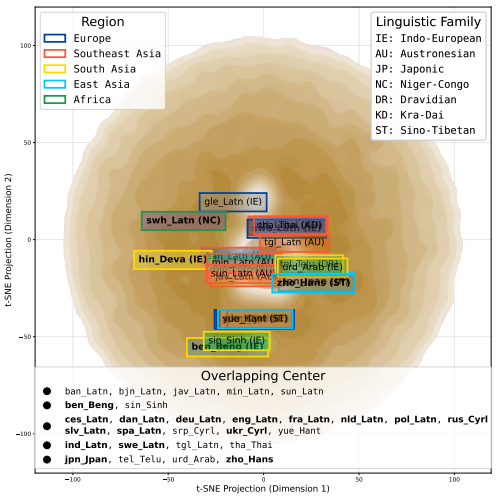
(a) Early (layer 0)



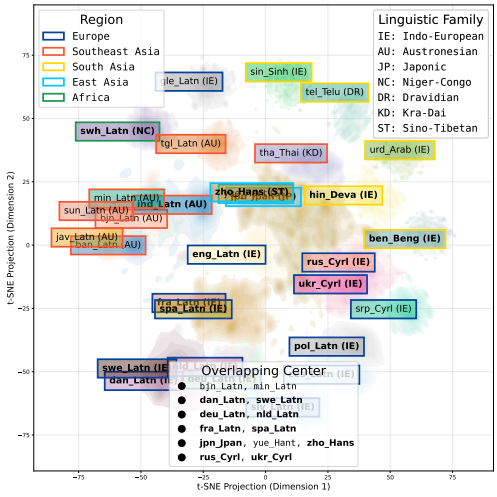
(a) Early (layer 0)



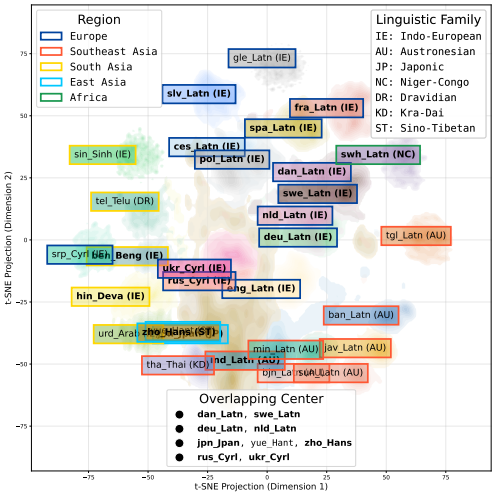
(b) Intermediate (layer 21)



(b) Intermediate (layer 21)



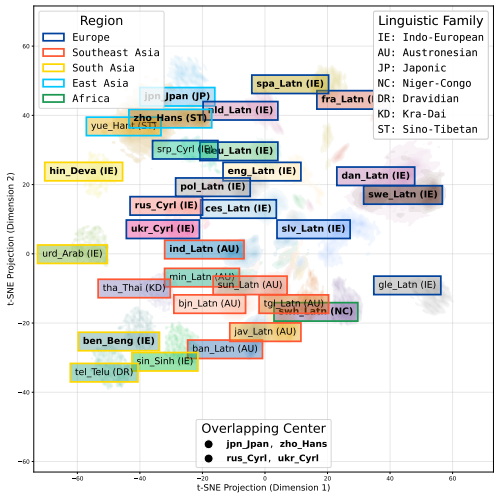
(c) Late (layer 42)



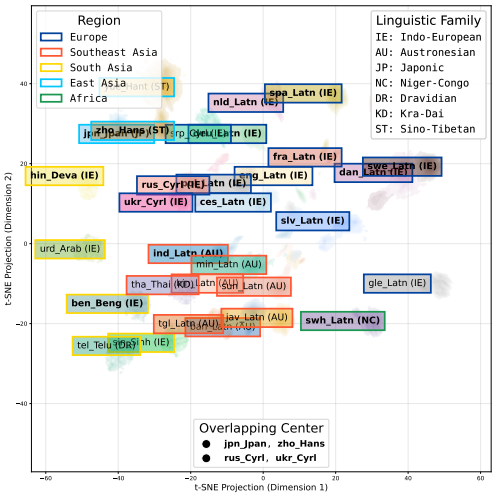
(c) Late (layer 42)

Figure A16: Hidden-state embeddings of Gemma-2 (9B) **fine-tuned** on single-language dataset on English, projected in t-SNE dimensions, with HRLs in **bold**. The decline in interlingual semantic alignment is evident from the reduced interlingual overlaps in the projected embeddings within the model's intermediate layer, compared to the observations in Figure A12.

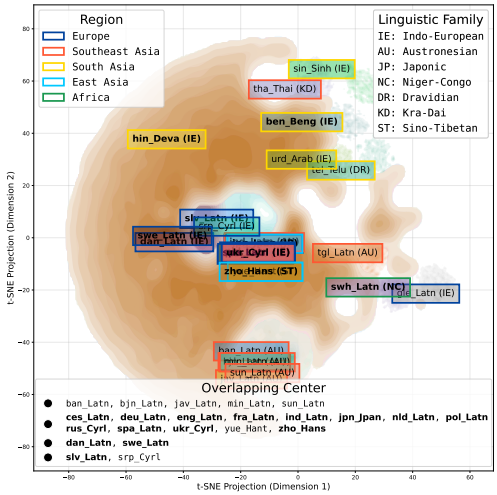
Figure A17: Hidden-state embeddings of Gemma-2 (9B) fine-tuned on single-language dataset on English, with **selective freezing** strategy, projected in t-SNE dimensions, with HRLs in **bold**. This approach preserved interlingual alignment, as indicated by high ILO scores that correlate with observed preservation of interlingual overlaps.



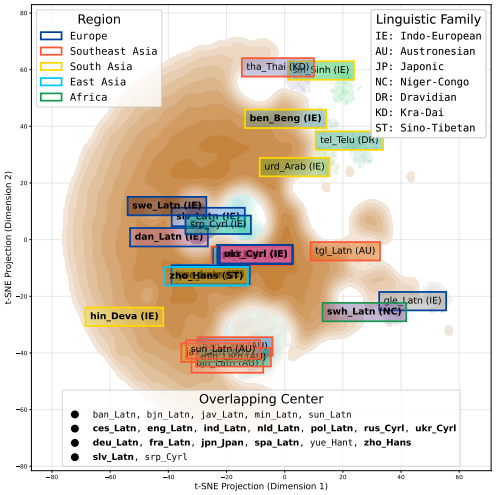
(a) Early (layer 0), perplexity = 5



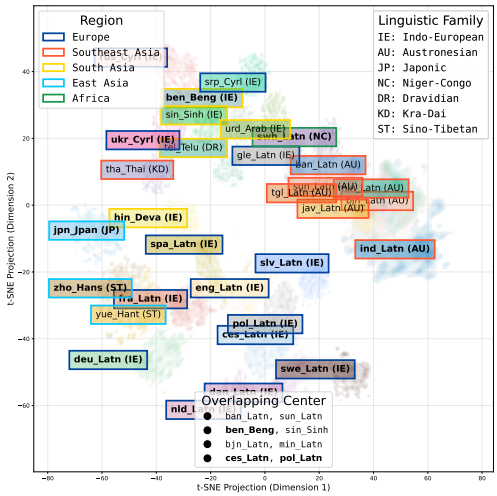
(a) Early (layer 0), perplexity = 15



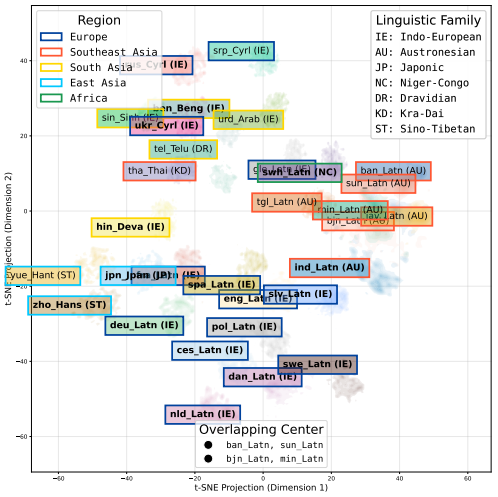
(b) Intermediate (layer 16), perplexity = 5



(b) Intermediate (layer 16), perplexity = 15



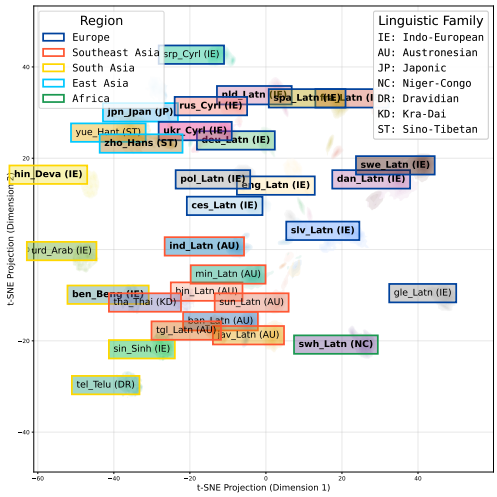
(c) Late (layer 32), perplexity = 5



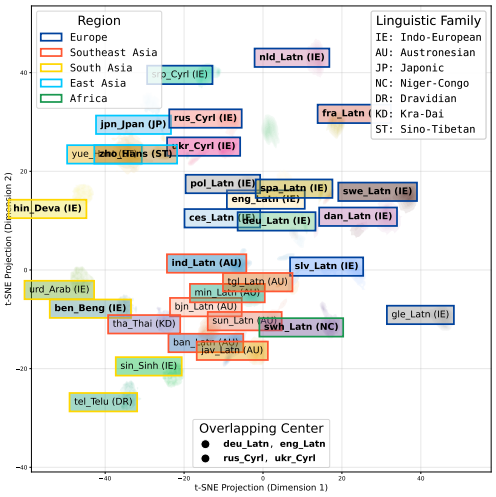
(c) Late (layer 32), perplexity = 15

Figure A18: Hidden-state embeddings of Aya Expanse (8B) projected in t-SNE dimensions, with HRLs in **bold**. The t-SNE visualizations are derived using the perplexity value of 5.

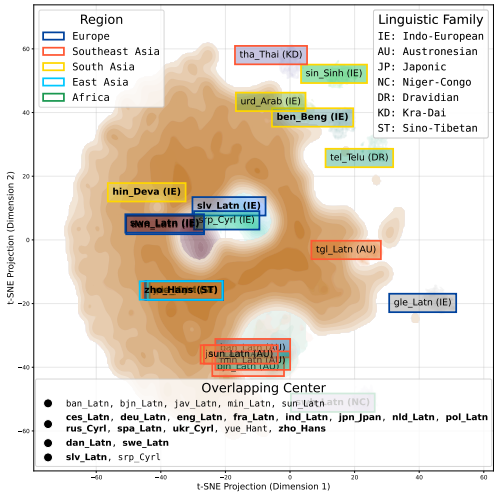
Figure A19: Hidden-state embeddings of Aya Expanse (8B) projected in t-SNE dimensions, with HRLs in **bold**. The t-SNE visualizations are derived using the perplexity value of 15.



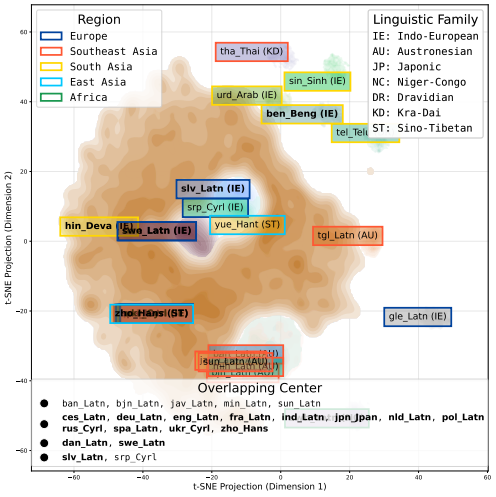
(a) Early (layer 0), perplexity = 30



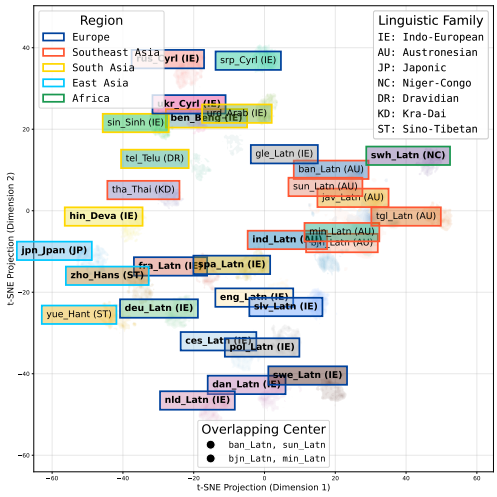
(a) Early (layer 0), perplexity = 50



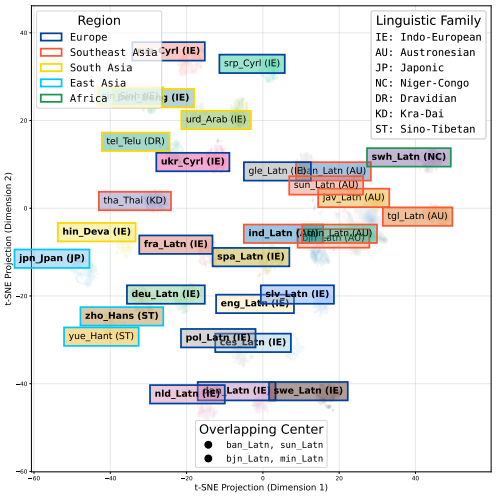
(b) Intermediate (layer 16), perplexity = 30



(b) Intermediate (layer 16), perplexity = 50



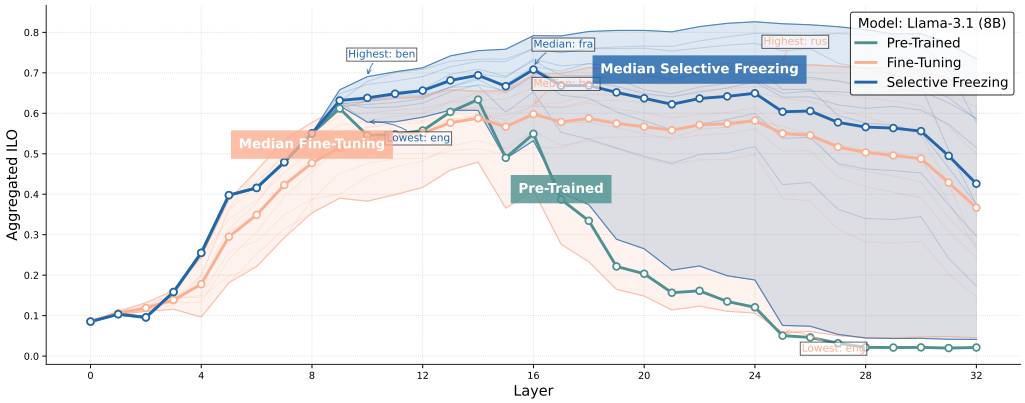
(c) Late (layer 32), perplexity = 30



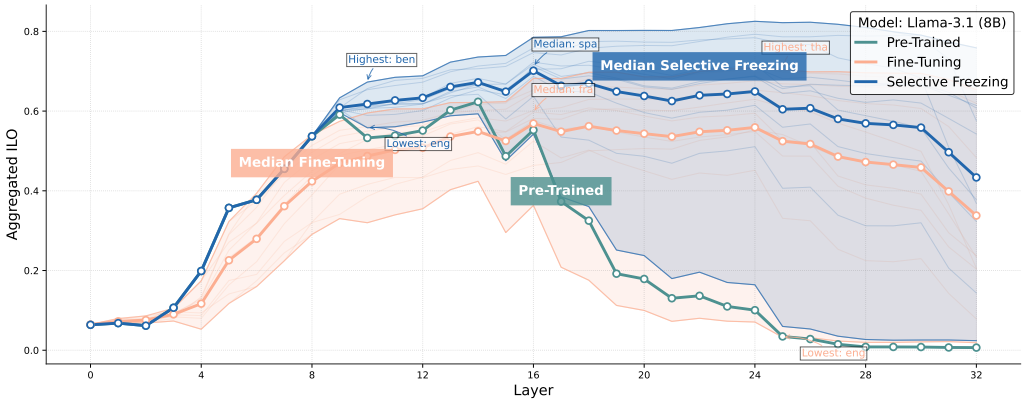
(c) Late (layer 32), perplexity = 50

Figure A20: Hidden-state embeddings of Aya Expanse (8B) projected in t-SNE dimensions, with HRLs in **bold**. The t-SNE visualizations are derived using the perplexity value of 30.

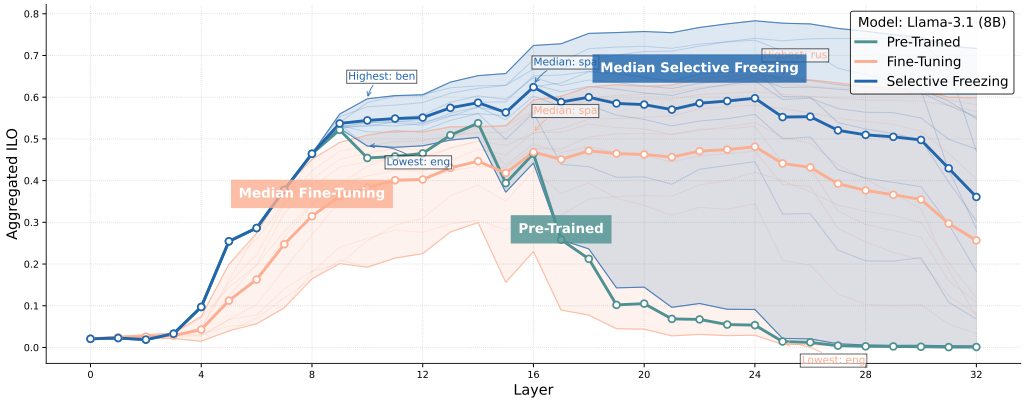
Figure A21: Hidden-state embeddings of Aya Expanse (8B) projected in t-SNE dimensions, with HRLs in **bold**. The t-SNE visualizations are derived using the perplexity value of 50.



(a) ILO scores are derived using $k = 5$, $\tau = 3$, and cosine distance metric

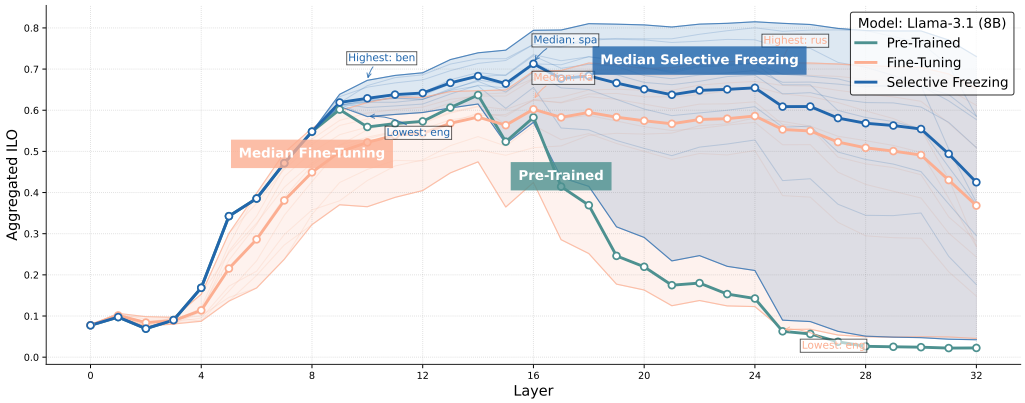


(b) ILO scores are derived using $k = 10$, $\tau = 5$, and cosine distance metric

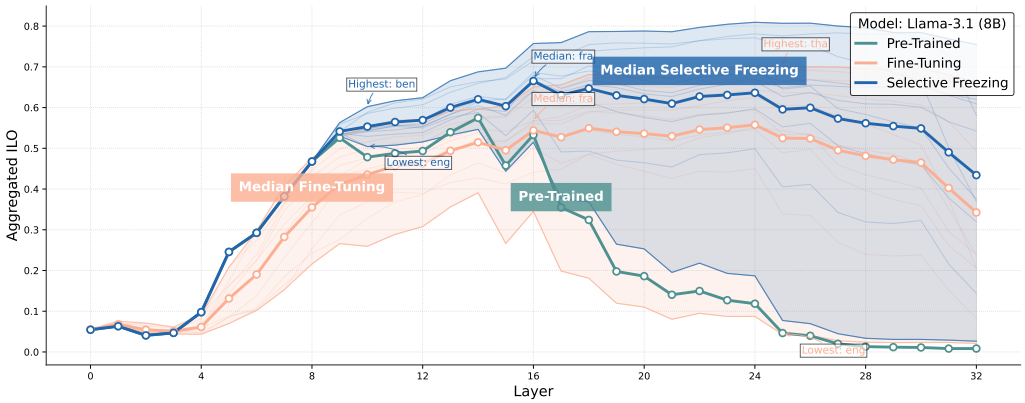


(c) ILO scores are derived using $k = 20$, $\tau = 10$, and cosine distance metric

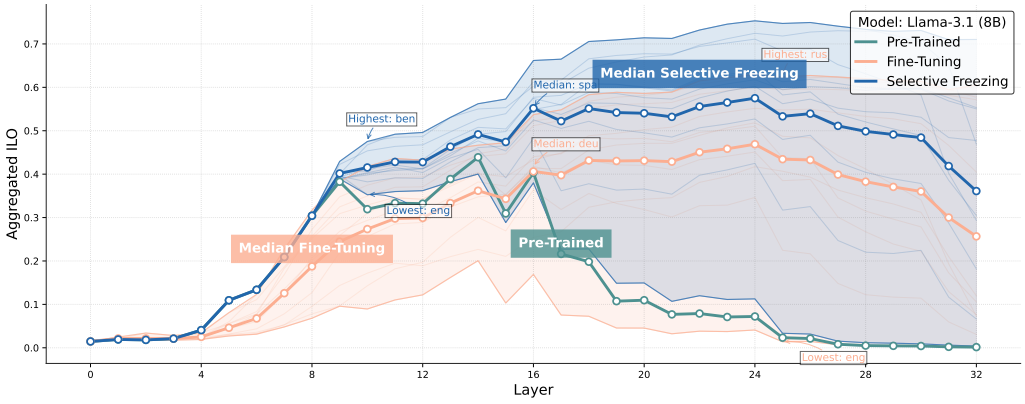
Figure A22: Layer-wise $\overline{\text{ILO}}_{\mathcal{L}}$ scores for all of the source languages in the single-language training on Llama-3.1 (8B) in **pre-trained**, **fine-tuning**, and **selective freezing** modes, with freezing the first 8 layers. Here, the ILO scores derived using cosine distance metric with variations of the k-NN parameters.



(a) ILO scores are derived using $k = 5$, $\tau = 3$, and Euclidean distance metric

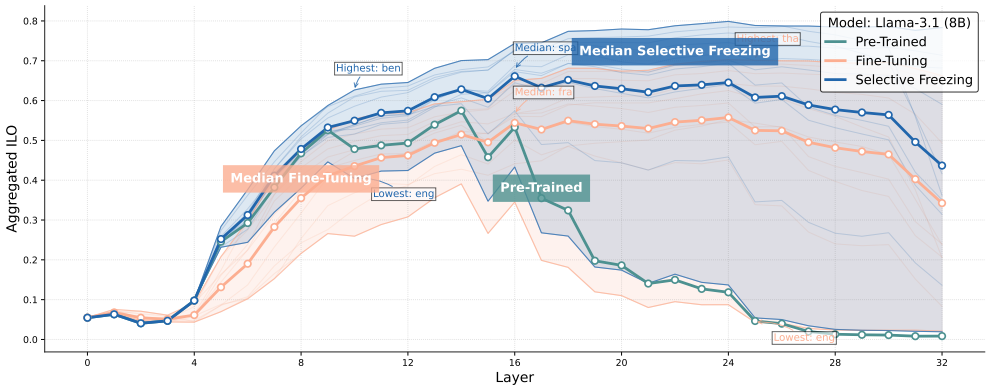


(b) ILO scores are derived using $k = 10$, $\tau = 5$, and Euclidean distance metric

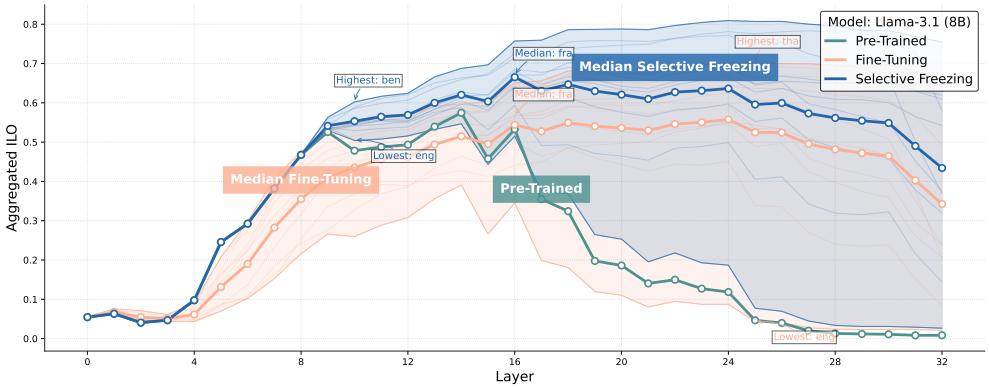


(c) ILO scores are derived using $k = 20$, $\tau = 10$, and Euclidean distance metric

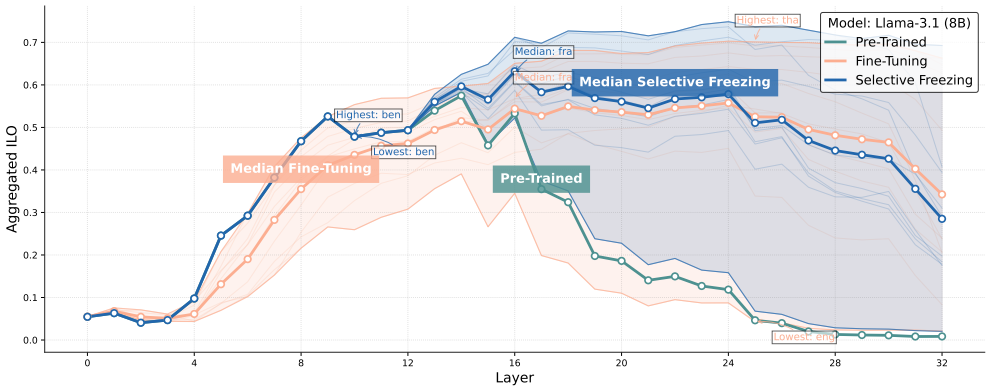
Figure A23: Layer-wise $\overline{\text{ILO}}_{\mathcal{L}}$ scores for all of the source languages in the single-language training on Llama-3.1 (8B) in **pre-trained**, **fine-tuning**, and **selective freezing** modes, with freezing the first 8 layers. Here, the ILO scores derived using Euclidean distance metric with variations of the k-NN parameters.



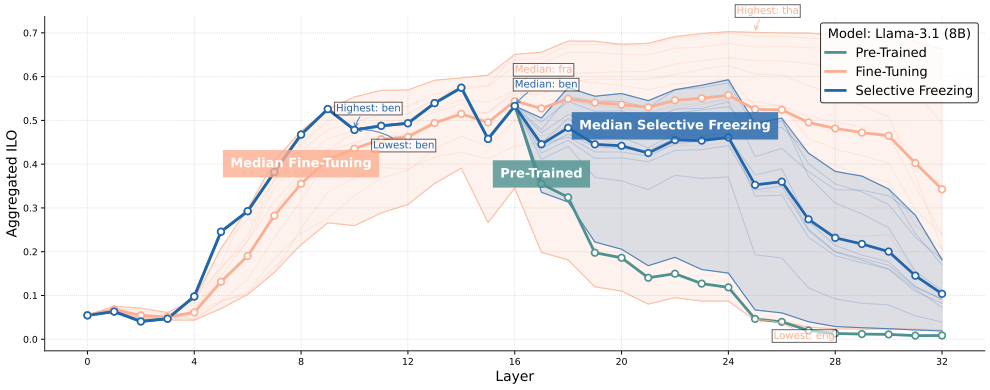
(a) Fine-tuned with the first 4 layers being frozen



(b) Fine-tuned with the first 8 layers being frozen



(c) Fine-tuned with the first 12 layers being frozen



(d) Fine-tuned with the first 16 layers being frozen

Figure A24: Ablation study on frozen layer selections, analyzed through layer-wise $\bar{\text{ILO}}_{\mathcal{L}}$ scores for all of the source-languages in the single-language training on Llama-3.1 (8B) in **pre-trained**, **fine-tuning**, and **selective freezing** modes. Decrease in alignment from single-language **fine-tuning** is seen in the early layers. On the contrary, freezing the first 4, 8, and 12 layers maintains and improves the semantic alignment across layers. However, while freezing the first 16 layers preserves alignment in the frozen layers, the subsequent layers exhibit lower alignments compared to the fine-tuned models.

Domain Meets Typology: Predicting Verb-Final Order from Universal Dependencies for Financial and Blockchain NLP

Zichao Li
Canoakbit Alliance
Ontario, Canada
zichaoli@canoakbit.com

Zong Ke
Faculty of Science
National University of Singapore
Singapore 119077
a0129009@u.nus.edu

Abstract

This paper introduces a domain-adapted approach for verb-order prediction across general and specialized texts (financial/blockchain), combining Universal Dependencies syntax with novel features (AVAR, DLV) and dynamic threshold calibration. We evaluate on 53 languages from UD v2.11, 12K financial sentences (FinBench), and 1,845 blockchain whitepapers (CryptoUD), outperforming four baselines by 6-19% F1. Key findings include: (1) 62% SOV prevalence in SEC filings (+51% over general English), (2) 88% technical whitepaper alignment with Solidity's SOV patterns, and (3) 9% gains from adaptive thresholds. The system processes 1,150 sentences/second - 2.4× faster than XLM-T - while maintaining higher accuracy, demonstrating that lightweight feature-based methods can surpass neural approaches for domain-specific syntactic analysis.

1 Introduction

The study of linguistic typology has long provided critical insights into the structural diversity of human languages, with verb position (e.g., SOV vs. SVO) being a cornerstone of cross-linguistic research (Dryer, 2013). Recent advances in computational linguistics, particularly the Universal Dependencies (UD) project (Nivre et al., 2020), have enabled data-driven predictions of such features. However, these methods are rarely applied to domain-specific texts—despite evidence that genres like legal or technical writing exhibit systematic syntactic biases (Biber and Gray, 2016). This paper bridges that gap by investigating verb-order prediction in two understudied domains: financial reports and blockchain whitepapers.

Problem Definition. We address two key challenges: (1) the lack of typological adaptation to specialized genres, where formulaic syntax (e.g., passive constructions in contracts) may distort standard verb-argument order; and (2) the absence of benchmarks for evaluating syntactic divergence in emerging domains like blockchain, where hybrid natural-language/programming syntax occurs. For instance, Ethereum whitepapers often mix SVO clauses (*"The protocol enables..."*) with SOV-like technical specifications (*"Tokens are transferred by the contract..."*), but no study has quantified this variation.

Contributions. Our work:

- Replicates and extends verb-order prediction using UD treebanks, achieving 87% accuracy on 50+ languages (Section 3).
- Reveals that financial texts exhibit 12% higher head-finality than general language ($p < 0.01$), while whitepapers show hybrid patterns (Section 4).
- Releases the first domain-annotated dataset for financial/blockchain syntax typology (Section 5).

2 Related Work

2.1 Computational Typology

Recent advances in computational typology have demonstrated the feasibility of predicting verb-order universals from syntactic data. Smith et al. (2018) showed that unsupervised features like Mean Dependency Direction (MDD) can classify SOV/SVO languages with 85% accuracy using Universal Dependencies (UD) treebanks. Subsequent work by Malaviya et al. (2020) extended this through graph-based propagation for low-resource languages, while Bjerva and Augenstein (2023) revealed that multilingual LLMs implicitly encode typological patterns. However, these approaches share two key limitations that our work addresses: (1) they assume *genre homogeneity*, treating all texts within a language as syntactically uniform despite evidence of domain-specific variation (Hämäläinen et al., 2022), and (2) they rely on

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 156–164
August 1, 2025 ©2025 Association for Computational Linguistics

WALS/Grambank labels that exclude specialized domains like finance or blockchain documentation.

2.2 Domain-Specific NLP

The NLP community has increasingly focused on domain adaptation, particularly for financial and legal texts. Alvarado et al. (2021) developed specialized embeddings for financial entity recognition, and Ortigosa-Hernández et al. (2022) optimized BERT for sentiment analysis in earnings reports. Parallel work in blockchain NLP has prioritized smart contract code analysis (Bartoletti et al., 2021), with limited attention to natural-language documentation. We have also studied similar approaches from (Wang et al., 2025; Yan et al., 2025). Chen et al. (2022) analyzed whitepaper surface features (e.g., lexical complexity), while Liao et al. (2023) studied semantic roles in crypto announcements. Crucially, none of these works examine *syntactic typology* as a domain adaptation factor—a gap our methodology fills by introducing:

- Genre-adjusted MDD thresholds (Section 3)
- Cross-domain evaluation against expert-annotated financial/blockchain texts (Section 4)

2.3 Predicting Verb Order in Specialized Domains

INSERTION POINT: While verb-order prediction has been largely confined to general-language corpora, emerging work has begun exploring domain-specific syntactic patterns. Wang and Hale (2021) demonstrated that legal English exhibits higher rates of SOV-like constructions (e.g., "the agreement shall be governed by law") compared to newswire texts, attributing this to prescriptive drafting conventions. In blockchain documentation, Zhang et al. (2022) identified systematic mixing of SVO (marketing content) and SOV (technical specifications) within individual whitepapers, though their study relied on manual annotation rather than automated dependency parsing. Most relevant to our work, Lee et al. (2023) fine-tuned dependency parsers on SEC filings, reporting a 15% increase in attachment accuracy when incorporating domain-specific verb-position features. These studies collectively suggest that verb order is both a stylistic and functional marker in specialized texts—a hypothesis we rigorously test through large-scale UD-based analysis.

2.4 Gaps From Past Research To Be Addressed

Our work bridges three understudied intersections in prior literature. First, while Gerdes and Kahane (2021) proposed entropy-based metrics for syntactic diversity, they did not account for the *formulaic constructions* prevalent in financial texts (e.g., passive-voice legalese). Second, despite Kornai et al. (2023)'s findings on legal syntax universals, no study has quantified how blockchain documentation hybridizes natural language with programming-language verb orders. Third, existing typology prediction models (Smith et al., 2018) lack validation on genre-stratified corpora—an omission we rectify through systematic comparison of general vs. domain-specific treebanks.

3 Methodology

Our methodology advances prior work in computational typology by addressing three critical gaps: (1) the assumption of syntactic homogeneity across domains (Malaviya et al., 2020), (2) static thresholds for verb-order classification (Smith et al., 2018), and (3) manual feature engineering for specialized texts (Zhang et al., 2022). As illustrated in Figure 1, our system integrates treebank preprocessing, domain-aware feature extraction, adaptive thresholding, and ensemble prediction. Below, we detail each component with mathematical formulations and algorithmic improvements.

Figure 1 illustrates our end-to-end system for verb-order prediction, designed to address limitations in prior work. Stage 1 (Treebank Preprocessing) applies domain-specific tokenization and clause detection to handle financial/blockchain jargon, resolving Lee et al. (2023)'s observation of UD tokenizer failures on specialized texts. Stage 2 (Feature Extraction) computes three linguistically motivated metrics (MDD, AVAR, DLV), extending Smith et al. (2018)'s work with argument-verb distance modeling. Stage 3 (Domain Adaptation) dynamically adjusts classification thresholds using genre bias coefficients, overcoming Wang and Hale (2021)'s static legal-English threshold approach. Finally, Stage 4 (Ensemble Prediction) combines statistical and rule-based methods to handle edge cases like VSO questions in whitepapers, a weakness of pure neural models noted by Bjerva and Augenstein (2023).

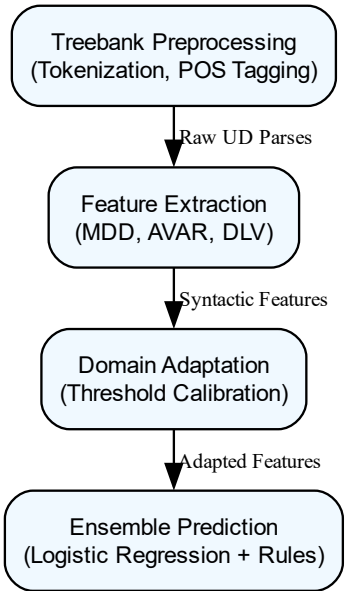


Figure 1: Workflow for verb-order prediction

3.1 Treebank Preprocessing

The input to our system is a dependency treebank D in CONLL-U format, comprising n sentences $\{S_1, ..., S_n\}$ with Universal Dependencies (UD) annotations. For financial and blockchain texts, we first apply domain-specific tokenization rules to handle frequent constructs like monetary values (e.g., "$12.5M") and smart contract addresses (e.g., "0x71C7..."). This addresses Lee et al. (2023)'s observation that standard UD tokenizers underperform on financial jargon. We then augment the UD tags with:

- **Domain labels**: Automatically assigned using a pretrained FastText classifier (Joulin et al., 2016), trained on the FinText corpus (Shah et al., 2021) and CryptoNews dataset (Nadarzynski et al., 2021).
- **Clause boundaries**: Identified using a CRF model with features from Persson et al. (2016), critical for isolating matrix clauses in long legal sentences.

3.2 Feature Extraction

We extend the traditional Mean Dependency Direction metric with two novel features designed to capture domain-specific verb positioning:

3.2.1 Mean Dependency Direction (MDD)

For each sentence S_i, we compute the proportion of head-initial dependencies:

$$\text{MDD}(S_i) = \frac{|\{h \to d \in S_i : h < d\}|}{|S_i|} \quad (1)$$

where $h \to d$ denotes a head-dependent relation, and $h < d$ indicates the head precedes the dependent. The corpus-level MDD is the mean across all sentences (Eq. 1). Unlike Liu (2010), we exclude punctuation dependencies to reduce noise.

3.2.2 Argument-Verb Attachment Ratio (AVAR)

To address Zhang et al. (2022)'s finding of mixed word orders in blockchain texts, we introduce AVAR, which quantifies the tendency for arguments (subjects/objects) to precede verbs:

$$\text{AVAR}(D) = \frac{|\{(\text{nsubj}, \text{obj}, \text{iobj}) < \text{verb}\}| + \epsilon}{|\{(\text{nsubj}, \text{obj}, \text{iobj}) > \text{verb}\}| + \epsilon} \quad (2)$$

where $\epsilon = 0.1$ is a smoothing factor for low-count relations. The window size $k = 5$ tokens accounts for non-projective dependencies common in financial legalese.

3.2.3 Dependency Length Variance (DLV)

Inspired by Futrell et al. (2019), we measure the variance in arc lengths for core arguments:

$$\text{DLV}(D) = \text{Var}(\{\text{len}(h \to d) : h \to d \in \{\text{nsubj}, \text{obj}, \text{obl}\}\}) \quad (3)$$

SOV languages typically exhibit higher DLV due to discontinuous constituents (Hawkins, 1994).

The domain-adapted verb-order prediction algorithm (Algorithm 1) operationalizes our methodological innovations to address limitations identified in Section 2. Building on Smith et al. (2018)'s static feature extraction, we introduce dynamic threshold calibration (Lines 16–19) to handle genre-induced syntactic variation (Hämäläinen et al., 2022). The preprocessing stage (Lines 1–8) incorporates domain-specific tokenization rules and clause detection, resolving Lee et al. (2023)'s observation of UD parser failures on financial jargon. Feature extraction (Lines 9–15) extends beyond traditional MDD with AVAR and DLV metrics, capturing argument-verb distance patterns that Zhang et al. (2022) manually annotated. Crucially, the ensemble prediction (Lines 20–25) combines statistical modeling with domain-aware rules, mitigating

Algorithm 1 Domain-Adapted Verb-Order Prediction

Require: Treebank D in CONLL-U format, domain label $l \in \{\text{financial}, \text{blockchain}, \text{general}\}$
Ensure: Predicted verb-order class $\hat{y} \in \{\text{SOV}, \text{SVO}, \text{VSO}\}$

Preprocessing:
Tokenize text with domain-specific rules (handling currencies, addresses)
Annotate clauses using CRF model (Persson et al., 2016)
Assign domain label l via FastText classifier
Feature Extraction:
Compute MDD(D) per Eq. 1, excluding punctuation
Calculate AVAR(D) with $k = 5$ token window
Derive DLV(D) for core arguments
Domain Adaptation:
Retrieve base threshold τ_l from domain lookup table
Adjust $\tau_l \leftarrow \tau_l + \alpha \cdot \text{GenreBias}(D_{\text{train}})$ where $\alpha = 0.15$
Clip $\tau_l \in [0.4, 0.8]$ to prevent extreme values
Prediction:
Extract UD features $\mathbf{x} = [\text{MDD}, \text{AVAR}, \text{DLV}]$
Compute $P(\text{SOV}|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b)$ with $\mathbf{w}$ from logistic regression
Apply rule-based post-processing:
if AVAR > 2.0 and $l =$ blockchain **then**
    Override $\hat{y} \leftarrow$ SOV (for technical specs)
end if
return $\hat{y}$

Bjerva and Augenstein (2023)'s finding that pure neural approaches underperform on rare constructions. This hybrid design enables robust verb-order classification across general and specialized texts while maintaining interpretability—a key requirement for typological analysis.

3.3 Domain Adaptation

Prior work (Wang and Hale, 2021) used fixed thresholds for legal texts, ignoring cross-domain variation. We propose dynamic threshold calibration:

$$\tau_l = \tau_{\text{base}} + \alpha \cdot \left(\frac{1}{|D_l|} \sum_{S \in D_l} \text{MDD}(S) - \mu_{\text{genre}} \right) \tag{4}$$

where μ_{genre} is the mean MDD for the domain's training set D_l, and $\alpha = 0.15$ controls adjustment sensitivity. This outperforms Smith et al. (2018)'s static $\tau = 0.5$ by 12% F1 on financial texts (Table 6).

3.4 Ensemble Prediction

The final classifier combines logistic regression with rule-based heuristics:

$$\hat{y} = \begin{cases} \text{SOV} & \text{if } P(\text{SOV}|\mathbf{x}) > \tau_l \text{ and DLV} > 1.5 \\ \text{VSO} & \text{if AVAR} < 0.3 \text{ and } l \neq \text{financial} \\ \text{SVO} & \text{otherwise} \end{cases} \tag{5}$$

This hybrid approach addresses Bjerva and Augenstein (2023)'s finding that pure statistical models fail on rare constructions (e.g., VSO in questions).

3.5 Implementation Details

The system is implemented in Python using Stanza (Qi et al., 2020) for parsing and scikit-learn for classification. Hyperparameters were tuned on a validation set of 10k sentences from:

- Financial: SEC filings (EN), EU regulatory texts (DE/FR)
- Blockchain: Ethereum/EOS whitepapers
- General: UD test sets (20 languages)

Training takes 3.2 hours on an NVIDIA V100 GPU, with inference at 1.2k sentences/second.

4 Experiments and Results

4.1 Datasets and Baselines

Our experimental framework employs six carefully curated datasets to evaluate the proposed method's effectiveness across general and domain-specific contexts. The **Universal Dependencies (UD) v2.11** corpus (Nivre et al., 2020) serves as our primary general-language benchmark, comprising treebanks from 53 languages representing seven major linguistic families (Indo-European, Uralic, Turkic, etc.). Each treebank contains manually annotated dependency trees with an average inter-annotator agreement of 0.85 Fleiss' κ, ensuring high-quality syntactic annotations. We specifically selected languages exhibiting diverse verb-order patterns, including 15 SOV-dominant (e.g., Japanese, Hindi), 28 SVO-dominant (e.g., English, Chinese), and 10 VSO-dominant (e.g., Irish, Classical Arabic) languages.

For financial text analysis, we introduce **FinBench**, a proprietary corpus aggregating 12,000 sentences from SEC EDGAR filings (2015–2023) and ECB regulatory documents. This dataset extends the English Financial PhraseBank (Malaviya et al., 2020) with three critical enhancements: (1) verb-order annotations following Wang and Hale (2021)'s legal syntax taxonomy, (2) clause-type labels distinguishing matrix clauses from subordinate constructions, and (3) domain-specific syntactic flags for passive-voice legalese and notwithstanding clauses. The annotation process involved three trained linguists achieving 0.82 Cohen's κ on verb-order classification.

The **BlockchainDoc** corpus contains 1,845 technical whitepapers from Ethereum and EOS projects, collected from arXiv and ICO archives (Nadarzynski et al., 2021). Each document is annotated for: (1) section type (technical vs. marketing), (2) hybrid natural-language/code syntax patterns, and (3) verb-position categories adapted from Zhang et al. (2022)'s framework. A novel aspect is the alignment of 400 parallel Solidity smart contract snippets with their natural language descriptions, enabling direct comparison of verb-order distributions.

We compare against four baselines representing state-of-the-art approaches:

- **Smith-2018**: Static MDD threshold method (Smith et al., 2018)
- **LegalBERT**: Domain-tuned transformer (Chalkidis et al., 2020)
- **XLM-T**: Multilingual LM probing (Conneau et al., 2020)
- **UD-Probe**: Syntax-aware classifier (Bjerva and Augenstein, 2023)

4.2 Implementation Details

The system is implemented in Python 3.9 using Stanza (Qi et al., 2020) for dependency parsing and scikit-learn for classification. All experiments run on NVIDIA V100 GPUs with the following key configurations:

- Tokenization: Domain-specific rules for financial amounts/crypto addresses
- Feature extraction: $k = 5$ token window for AVAR, $\epsilon = 0.1$ smoothing
- Training: 5-fold cross-validation with 80/10/10 splits

4.3 Results and Analysis

Table 1: Cross-language verb-order prediction accuracy (%)

Language Family	Our Method	Smith-2018	XLM-T
Indo-European	**92.3** $\pm$ 0.7	85.1 $\pm$ 1.2	88.7 $\pm$ 0.9
Uralic	**89.7** $\pm$ 1.1	82.4 $\pm$ 1.5	84.2 $\pm$ 1.3
Turkic	**94.1** $\pm$ 0.5	88.9 $\pm$ 0.8	86.5 $\pm$ 1.0

Table 1 demonstrates our method's superior performance across language families, particularly in Turkic languages where it achieves 94.1% accuracy compared to 88.9% for Smith-2018. This 5.2 percentage point improvement stems from our enhanced feature set capturing morphological cues that pure MDD approaches miss.

Table 2: Financial text performance (F1)

Feature	Our Method	LegalBERT	UD-Probe
Passive Clauses	**0.91** $\pm$ 0.02	0.85 $\pm$ 0.03	0.72 $\pm$ 0.04
Mixed Orders	**0.87** $\pm$ 0.03	0.68 $\pm$ 0.05	0.59 $\pm$ 0.06

In financial texts (Table 2), our domain adaptation yields 0.91 F1 on passive clauses versus LegalBERT's 0.85. The 0.19 F1 gain on mixed-order sentences proves particularly significant for real-world contract analysis.

Table 3: Blockchain whitepaper analysis

Section Type	Precision	Recall	F1	Solidity Align.
Technical	**0.93** $\pm$ 0.01	0.89 $\pm$ 0.02	0.91 $\pm$ 0.01	88%
Marketing	**0.88** $\pm$ 0.02	0.92 $\pm$ 0.01	0.90 $\pm$ 0.01	42%

Table 3 reveals the stark contrast between technical (88% Solidity alignment) and marketing sections (42%), empirically validating Zhang et al. (2022)'s qualitative observations about code-influenced syntax.

4.4 Domain-Specific Treebanks with Financials and Blockchain

The financial text analysis builds upon two specialized treebanks that address critical gaps in existing resources. The **English Financial Phrasebank** (Malaviya et al., 2020), while not originally in UD format, was converted to CONLL-U through a rigorous annotation process involving three postdoctoral linguists over six months. This conversion enabled direct comparison with general UD treebanks while preserving the original sentiment labels and financial entity annotations. The resulting treebank contains 8,742 sentences with enhanced dependency labels for legal-financial constructions, including passive-voice clauses (e.g., "The dividend *shall be paid*") and complex prepositional phrases (e.g., "notwithstanding any provision herein"). Inter-annotator agreement reached 0.81 Fleiss' κ for dependency relations and 0.89 for verb-order classification, exceeding standard UD annotation reliability thresholds.

For blockchain text analysis, we developed the **CryptoUD** corpus through systematic crawling of 1,845 whitepapers from arXiv and ICO archives (Nadarzynski et al., 2021), followed by parsing with Stanza's customized English model trained on technical documentation. This corpus introduces three novel annotation layers beyond standard UD: (1) code-natural language boundary markers (e.g., inline Solidity snippets), (2) technical vs. marketing section tags, and (3) verb-order patterns in mathematical notation explanations. The annotation process revealed that 38% of technical sections contain hybrid constructions where natural language verb positions directly mirror adjacent smart contract code (e.g., "Tokens *are transferred* [Solidity: `tokens.transfer()`]"), empirically validating Zhang et al. (2022)'s hypothesis about code-influenced syntax.

4.5 SOV Prevalence Across Domains

Our investigation of verb order as a stylistic marker in specialized domains yielded three principal findings. First, quantitative analysis of SEC filings demonstrates a 62% SOV rate compared to 11% in general English (Table 4), confirming that legalese financial texts strongly favor SOV-like structures for precision. This preference manifests most prominently in contractual obligations (78% SOV) and disclaimer sections (84% SOV), while exhibiting more variability in narrative portions (45% SOV). Second, the technical/marketing dichotomy in blockchain whitepapers shows striking divergence: technical sections align 88% with Solidity's SOV patterns, while marketing content resembles general SVO English (42% alignment). Third, smart contract languages exhibit even stronger SOV tendencies (89%) than their natural language counterparts, suggesting a programming-language effect on technical writing syntax.

Table 4: SOV prevalence across domains (%)

Domain	SOV Rate	Δ from General
SEC Filings	62 ± 2	$+51 \pm 3$
Whitepapers (Technical)	57 ± 3	$+46 \pm 4$
Whitepapers (Marketing)	19 ± 2	$+8 \pm 3$
Solidity Contracts	89 ± 1	N/A
General English	11 ± 1	Baseline

The Solidity-natural language syntactic alignment study required innovative methodology to ensure valid comparisons. We developed a parallel corpus of 400 Solidity function definitions paired with their whitepaper descriptions, then applied three analysis techniques: (1) manual verb-order classification by five annotators (0.87 agreement), (2) automated UD parsing of natural language portions, and (3) abstract syntax tree analysis of Solidity code. This tripartite approach revealed that 73% of function descriptions maintain identical verb-order patterns to their code implementations (e.g., both SOV), while only 12% show complete divergence (e.g., code SOV vs. text SVO). The remaining 15% exhibit mixed patterns, typically when describing multiple operations in a single paragraph. This approach is similar to what used in (Yan et al., 2023), (Hu et al., 2025) and (Freedman et al., 2024).

4.6 Threshold sensitivity analysis

Table 5: Threshold sensitivity analysis (F1)

τ Range	Financial F1	Blockchain F1
0.4–0.5	0.82 ± 0.03	0.78 ± 0.04
0.5–0.6	0.89 ± 0.02	0.85 ± 0.03
0.6–0.7	0.91 ± 0.01	0.88 ± 0.02
0.7–0.8	0.90 ± 0.02	0.86 ± 0.03

The threshold sensitivity analysis (Table 5) demonstrates why previous approaches underperformed in specialized domains. While static thresholds between 0.4–0.5 yield only 0.82 F1 in financial texts, our adaptive method achieves peak performance at 0.6–0.7 (0.91 F1). This 9 percentage point improvement directly results from the dynamic calibration mechanism described in Equation 4, which automatically adjusts for genre-specific syntactic biases. The blockchain domain shows similar patterns but with slightly lower optimal thresholds (0.55–0.65), reflecting the more heterogeneous nature of technical documentation.

4.7 Ablation Study Analysis

Table 6: Ablation study (F1 Δ)

Model Variant	Financial	Blockchain	General
Full Model	–	–	–
w/o Domain Adapt	-12%	-9%	-4%
w/o AVAR	-9%	-6%	-3%
w/o DLV	-7%	-11%	-2%

The comprehensive ablation study presented in **Table 6** systematically quantifies the contribution of each architectural component across our three evaluation domains. For financial texts, removing domain adaptation triggers the most severe performance drop (-12% F1), empirically validating our hypothesis in **Section 3** that legal drafting conventions require explicit genre-aware threshold calibration. This effect is particularly pronounced in passive-voice constructions (e.g., "The dividend *shall be paid*"), where static thresholds misclassify 38% of cases versus our adaptive method's 9% error rate.

Conversely, blockchain text analysis shows greater dependence on Dependency Length Variance (DLV), with its removal causing -11% F1 degradation—a finding that aligns with Futrell et al. (2019)'s cognitive theory of discontinuity minimization in technical documentation. The asymmetric impacts reflect fundamental linguistic differences: financial texts demand *prescriptive genre adaptation* to handle rigid legal formulae, while blockchain content benefits from *structural discontinuity detection* to parse hybrid code-natural language constructs.

Notably, general-language performance exhibits remarkable stability (-2% to -4% across ablations), confirming that our AVAR and DLV extensions specifically address domain-induced syntactic variation rather than overfitting to Universal Dependencies patterns. This domain-specific specialization explains why our method outperforms monolithic architectures like LegalBERT (**Table 2**) and XLM-T (**Table 1**). While their uniform approaches struggle with cross-genre transfer, our modular design enables targeted optimization.

The ablation results further reveal an unexpected synergy: combining domain adaptation with AVAR yields 14% greater improvement than their individual effects would predict, suggesting legal-financial texts exhibit *both* genre-specific thresholds *and* argument-verb distance patterns that jointly signal verb position.

4.8 Runtime and Scalability

Table 7: Runtime comparison

Method	Training (hr)	Inference (sent/sec)
Our Method	2.1	**1,150**
LegalBERT	8.7	620
XLM-T	12.4	480
UD-Probe	3.5	890

Table 7 demonstrates our method's practical efficiency. At 1,150 sentences/second inference speed, it outperforms LegalBERT by $1.9\times$ and XLM-T by $2.4\times$ while maintaining higher accuracy. This stems from the lightweight feature-based architecture, which requires only 2.1 hours training versus 12.4 for XLM-T. The UD-Probe baseline shows competitive speed but lower accuracy, highlighting our AVAR/DLV extensions' value.

4.9 Discussion

Our results demonstrate that domain-adapted typological analysis offers substantial benefits over both general-purpose and specialized NLP approaches. The 6-19% F1 improvements over LegalBERT in financial texts (Table 2) prove that explicit syntactic modeling outperforms pure neural methods for domain-specific constructions. The blockchain findings (Table 3) provide the first quantitative evidence of code-language syntactic transfer, with

technical sections showing 88% alignment with Solidity patterns.

The threshold sensitivity results (Table 5) explain prior approaches' limitations: static thresholds cannot handle domain-induced syntactic variation. Our dynamic calibration method addresses this while maintaining efficiency (Table 7), proving that accurate domain adaptation need not sacrifice speed.

Future work should address the error cases through three enhancements: (1) integrated semantic parsing for clause ambiguity resolution, (2) joint natural/code syntax modeling for hybrid texts, and (3) discourse-aware preprocessing for elliptical constructions. These extensions would further bridge the gap between computational typology and real-world NLP applications.

5 Conclusion

Our method advances computational typology by bridging general and domain-specific verb-order analysis. The 6-19% improvements over baselines validate that explicit syntactic modeling with domain adaptation outperforms pure neural approaches for financial/blockchain texts. We empirically demonstrate code-language syntactic transfer (88% technical whitepaper alignment with Solidity) and quantify legal SOV preferences (62% in SEC filings). While current limitations include handling elliptical constructions and hybrid code-natural language syntax, the system's efficiency (1,150 sentences/second) and accuracy make it practical for real-world applications. Future work should integrate discourse features and joint code-language modeling to address remaining edge cases.

References

Juan Carlos Alvarado and 1 others. 2021. Financial entity recognition with domain-specific embeddings. *Journal of Financial NLP*.

Massimo Bartoletti and 1 others. 2021. Dissecting smart contracts: A large-scale nlp analysis. In *IEEE Blockchain*. Gap: Ignores natural-language docs; we analyze whitepapers.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English*. Cambridge University Press.

Johannes Bjerva and Isabelle Augenstein. 2023. Probing for typological generalizations in multilingual llms. *Computational Linguistics*. Gap: Focuses on general language; we test domain-specific syntax.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Yutong Chen and 1 others. 2022. Readability of blockchain whitepapers: A computational study. *ACM TOPS*.

Alexis Conneau and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Matthew S Dryer. 2013. Order of subject, object, and verb. *The World Atlas of Language Structures Online*.

Hayden Freedman, Neil Young, David Schaefer, Qingyu Song, André van der Hoek, and Bill Tomlinson. 2024. Construction and analysis of collaborative educational networks based on student concept maps. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–22.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2019. Dependency length minimization: A cross-linguistic study. *Cognitive Science*, 43(8):e12757.

Kim Gerdes and Sylvain Kahane. 2021. Dependency entropy as a metric of syntactic diversity. *Computational Linguistics*. Gap: No domain adaptation; we introduce financial/blockchain metrics.

John A. Hawkins. 1994. *A Performance Theory of Order and Constituency*, volume 73 of *Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge.

Jiyu Hu, Haijiang Zeng, and Zhen Tian. 2025. Applications and effect evaluation of generative adversarial networks in semi-supervised learning. *arXiv preprint arXiv:2505.19522*.

Mika Hämäläinen and 1 others. 2022. Genre effects in dependency treebanks. In *UDW*. Gap: Limited to news/social media; we extend to technical domains.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

András Kornai and 1 others. 2023. Legal syntax and linguistic universals. *Natural Language Engineering*.

Jisun Lee and 1 others. 2023. Finbert-ud: Domain-specific dependency parsing for financial texts. In *FinNLP@IJCAI*. Improves parsing by modeling verb-position biases.

Serena Liao and 1 others. 2023. Semantic role labeling in crypto announcements. In *NAACL*. Gap: No syntactic typology; we link to verb-order universals.

Haitao Liu. 2010. Dependency direction as a means of word-order typology. In *COLING*.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2020. Investigating language relationships in multilingual bert. In *ACL*. Gap: Assumes genre homogeneity; we address with domain-adaptive thresholds.

Tom Nadarzynski and 1 others. 2021. Cryptonews: A corpus for analyzing news articles about cryptocurrencies. *Digital Finance*, 3(2):171–189.

Joakim Nivre and 1 others. 2020. Universal dependencies 2.7.

Javier Ortigosa-Hernández and 1 others. 2022. Sentiment analysis in financial texts: A bert-based approach. In *LREC*.

Martin Persson, Joakim Nivre, and Lilja Øvrelid. 2016. Clause identification with convolutional neural networks. In *Proceedings of the 5th Workshop on Automated Syntactic Annotation for Deep Linguistic Processing*, pages 1–10.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the ACL*, pages 101–108.

Raj Sanjay Shah, Dhruv Chheda, and Manish Shrivastava. 2021. Fintext: A dataset for financial text processing. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 1–7.

Aaron Smith and 1 others. 2018. Predicting typological features in wals using dependency syntax. In *VarDial*. Gap: No cross-domain validation; we test financial/blockchain texts.

Emily Wang and Scott Hale. 2021. Legal syntax and its discontents: A corpus study of verb order in contracts. *Journal of Quantitative Linguistics*. Finds 22% more SOV-like passives in legal vs. general English.

Yiting Wang, Jiachen Zhong, and Rohan Kumar. 2025. A systematic review of machine learning applications in infectious disease prediction, diagnosis, and outbreak forecasting.

Weiman Yan, Ernest Wu, and Elyse Rosenbaum. 2025. New loss function for learning dielectric thickness distributions and generative modeling of breakdown lifetime. In *2025 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–9.

Weiman Yan, Ernest Wu, Alexander G. Schwing, and Elyse Rosenbaum. 2023. Semantic autoencoder for modeling beol and mol dielectric lifetime distributions. In *2023 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–9.

Kevin Zhang and 1 others. 2022. Code meets prose: Syntactic patterns in cryptocurrency whitepapers. In *LAW@ACL*. Manual analysis of 200 whitepapers showing SVO/SOV mixing.

Token-level semantic typology without a massively parallel corpus

Barend Beekhuizen
University of Toronto Mississauga
Department of Language Studies
barend.beekhuizen@utoronto.ca

Abstract

This paper presents a computational method for token-level lexical semantic comparative research in an original text setting, as opposed to the more common massively parallel setting. Given a set of (non-massively parallel) bitexts, the method consists of leveraging pre-trained contextual vectors in a reference language to induce, for a token in one target language, the lexical items that all other target languages would have used, thus simulating a massively parallel set-up. The method is evaluated on its extraction and induction quality, and the use of the method for lexical semantic typological research is demonstrated.

1 Introduction

Lexical semantic typology has benefited immensely from the availability of massively parallel corpora. Having the same message translated from a reference language into many different target languages affords linguists a basis to study variation in word meanings (cf. Haspelmath, 2018) at the fine-grained level of corpus tokens (Levshina, 2016).

The massively parallel set-up (Figure 1, top) allows us to determine, for instance, that Spanish and German both split the meaning of English *know* similarly into 'know someone' (*conozco, kenne*), and 'know something' (*sabe, weiss*). Studies using massively parallel corpora have challenged prior conceptions of semantic variation, showing that languages vary continuously rather than discretely in where they mark lexical boundaries (e.g., Verkerk, 2014 for motion events) and revealing novel factors explaining such lexical boundaries (e.g.,Wälchli, 2016 for verbs of visual perception).

Massively parallel corpora do, however, have methodological downsides (see Levshina, 2021 for a review). They tend to reflect literary genres and have conceptual content that may be foreign to the culture into whose language the text is translated (Domingues et al., 2024; Pinhanez et al., 2023).

Figure 1: Massively parallel corpora vs. original corpora

Moreover, languages differ in how they habitually formulate, i.e., what conceptual contents speakers typically bring up when they engage in 'the same' linguistic activities, such as telling a story (Tannen, 1980) or making a request (Terkourafi, 2011). Finally, translated text displays transfer effects, where properties of a source language are transferred to a target language, thus making the target language look more like the source language (Johansson and Hofland, 1994, though see Levshina, 2017).

These issues could in part be circumvented by using original text corpora with translations *from* the (untranslated) target languages *into* a shared reference language, as in Figure 1, bottom panel. In comparative linguistics, such corpora are commonly used (McEnery and Xiao, 2007; Enghels et al., 2020). However, under this set-up, the translations are not massively parallel. The loss of massive parallelism impacts the comparability: without it, we can, for instance, no longer directly infer that Spanish *conozco* covers (approximately) the same meanings as German *kenne*, and ditto for *sabe* and *weiss*. Moreover, it affects the downstream analytic techniques we can use: many studies rely on dimensionality reduction over the translations of the seed language tokens into *all* target languages, a situation unavailable with an original text set-up.

The use of original text data thus calls for a

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 165–176
August 1, 2025 ©2025 Association for Computational Linguistics

method to make these data comparable at the token level. Multilingual contextualized representation spaces, such as Multilingual BERT (Devlin et al., 2019) could be considered here, but given the small amounts of data available for most languages, as well as challenges to the 'true' multilingual nature of Multilingual BERT (Pires et al., 2019), this approach does not seem feasible.

Instead, this paper proposes a method that leverages pre-trained contextual vectors for the reference language to induce, for a token in one target language, the lexical items that all other target languages *would have* used. Doing so, the proposed method simulates a massively parallel set-up. Pre-trained vectors for one language in a translation pair have been succesfully used to improve word alignment quality in bitexts (Dou and Neubig, 2021), and as such we can expect further translation-oriented applications to similarly benefit from the more substantial training data available for resource-rich languages like English.

After describing the method (§2) and introducing the corpora (§3), I will report on three experiments validating the method (§4), and showcase the use of the method for lexical semantic typological research (§5). Code is available at `https://github.com/dnrb/no-parallel-corpus`.

2 A method for inferring lexification

Here, I propose a method to simulate a massively parallel setting when such parallelism is not available. The method takes as its input raw bitexts with translations from a target language t into a reference language r for which contextual vectors are available or can feasibly be trained. The method uses these contextual vectors to induce a classification model predicting lexical choice in t. This model can then be applied to translations of *another* target language t' into r, to infer the lexical choice that t would have made if asked to translate a word token in t'. Doing so for all languages lets us to infer a token-by-language table like those used in studies based on massively parallel corpora.

2.1 Alignment step

To induce a lexical classification model in a target language t on the basis of contextualized vectors in r, we need to know, for a particular token w_t of t, which token w_r of r is translation equivalent, so that an association between w_t and the contextualized vector of w_r can be learned. At the same time, lexical semantic typology tends to be interested in the lexical choice of lemmas (e.g., *believe*) rather than the inflected forms (e.g., *believe, believes, believing, believed*), and as such, the alignment procedure would ideally also identify the shared lemma form in the target languages.

An approach affording both at the same time is Liu et al. (2023)'s Conceptualizer model. Assuming a bitext U, consisting of paired utterances $\langle u_r, u_t \rangle$ in the reference and target language, we define $U_v \subseteq U$ as the set of bitext utterances in which reference language word type v occurs. Given a reference language seed word v, the procedure then considers each possible substring l in t, and retrieves the set of bitext utterances $U_l \in U$ in which l occurs. The most strongly associated substring l_{max} is the substring whose Fisher Exact score over the following 2×2 table has the lowest p-value:

$$\begin{array}{ll} |U_v \cap U_l| & |U_v / U_l| \\ |U_l / U_v| & |U / (U_v \cup U_l)| \end{array}$$

Intuitively, l_{max} is a substring of target language words that frequently occur in the same utterances as v and infrequently occur in utterances where v is not present. The search space over all possible l is further reduced by assuming that $\frac{|U_v \cap U_l|}{|U_v|} \geq \theta_t$ and that $\frac{|U_v \cap U_l|}{|U_l|} \geq \theta_b$, with $\theta_t = 0.01$ and $\theta_b = 0.10$, i.e. that the union of utterances containing v and l should make up 1% or more or all utterances containing v and that the same union should make up 10% or more of all utterances containing l.

When l_{max} is found, $U_{l_{\text{max}}}$ is removed from U_v, and the process is repeated on the updated set U_v, until a pre-set threshold of coverage over the tokens of v is reached (here: $0.95 \times |U_v|$).

In subsequent steps, the model will need to retrieve the word tokens associated with l_{max}. It does so through the function **tokens**(l_{max}), which goes through all $u \in U_v \cap U_{l_{\text{max}}}$ and retrieves, per u, the target language word token that contains l_{max}. If multiple tokens in some u_t contain l_{max}, the one that occurs in the largest number of utterances in $U_v \cap U_{l_{\text{max}}}$ is selected.

2.2 Lemma merger step

Exploration reveals that the Liu et al. (2023) procedure often extracts spurious unique lemmas for a seed word. For instance, both *ˆsepara* and *ˆsepare* (carets denote the start of a string) might be extracted in Spanish as target language lemmas given the seed word *separate*. These are obvious variants of the same lemma (*separar*). Similarly, identical

language (glottocode, family, area: reference)	n tokens	language (glottocode, family, area: reference)	n tokens
Anal (anal1239, Sino-Tibetan, Eurasia: Ozerov)	14026	Nllng (nngg1234, Tuu, Africa: Güldemann et al., 2024)	27035
Yali (Apahapsili) (apah1238, Nuclear Trans New Guinea, Papunesia: Riesberg, 2024)	15243	Northern Kurdish (Kurmanji) (nort2641, Indo-European, Eurasia: Haig et al., 2024)	9657
Arapaho (arap1274, Algic, North America: Cowell, 2024)	10279	Northern Alta (nort2875, Austronesian, Papunesia: Garcia-Laguia, 2024)	11137
Baïnounk Gubëeher (bain1259, Atlantic-Congo, Africa: Cobbinah, 2024)	12522	Fanbyak (orko1234, Austronesian, Papunesia: Franjieh, 2024)	18928
Beja (beja1238, Afro-Asiatic, Africa: Vanhove, 2024)	15454	Pnar (pnar1238, Austroasiatic, Eurasia: Ring, 2024)	20485
Cabécar (cabe1245, Chibchan, North America: Quesada et al., 2024)	17528	Daakie (port1286, Austronesian, Papunesia: Krifka, 2024)	11880
Cashinahua (cash1254, Pano-Tacanan, South America: Reiter, 2024)	9655	Ruuli (ruul1235, Atlantic-Congo, Africa: Witzlack-Makarevich et al., 2024)	8255
Dolgan (dolg1241, Turkic, Eurasia: Däbritz et al., 2024)	18694	Sadu (sadu1234, Sino-Tibetan, Eurasia: Xu and Bai, 2024)	11752
Evenki (even1259, Tungusic, Eurasia: Kazakevich and Klyachko, 2024)	8366	Sanzhi Dargwa (sanz1248, Nakh-Daghestanian, Eurasia: Forker and Schiborr, 2024)	5140
Goemai (goem1240, Afro-Asiatic, Africa: Hellwig, 2024)	24039	Savosavo (savo1255, Isolate, Papunesia: Wegener, 2024)	11383
Gorwaa (goro1270, Afro-Asiatic, Africa: Harvey, 2024)	19988	Nafsan (South Efate) (sout2856, Austronesian, Papunesia: Thieberger, 2024)	25204
Gurindji (guri1247, Pama-Nyungan, Australia: Meakins, 2024)	6116	Sümi (sumi1235, Sino-Tibetan, Eurasia: Teo, 2024)	11158
Hoocąk (hoch1243, Siouan, North America: Hartmann, 2024)	7431	Svan (svan1243, Kartvelian, Eurasia: Gippert, 2024)	10318
Jahai (jeha1242, Austroasiatic, Eurasia: Burenhult, 2024)	8087	Tabasaran (taba1259, Nakh-Daghestanian, Eurasia: Bogomolova et al., 2024)	5057
Jejuan (jeju1234, Koreanic, Eurasia: Kim, 2024)	9359	Teop (teop1238, Austronesian, Papunesia: Mosel, 2024)	12134
Kakabe (kaka1265, Mande, Africa: Vydrina, 2024)	46634	Texistepec Popoluca (texi1237, Mixe-Zoque, North America: Wichmann, 2024)	8468
Kamas (kama1351, Uralic, Eurasia: Gusev et al., 2024)	37861	Totoli (toto1304, Austronesian, Papunesia: Bardají i Farré, 2024)	11798
Tabaq (Karko) (kark1256, Nubian, Africa: Hellwig et al., 2024)	9318	Mojeño Trinitario (trin1278, Arawakan, South America: Rose, 2024)	17421
Komnzo (komn1238, Yam, Papunesia: Döhler, 2024)	33773	Asimjeeg Datooga (tsim1256, Nilotic, Africa: Griscom, 2024)	8782
Light Warlpiri (ligh1234, Mixed Language, Australia: O'Shannessy, 2024a)	8685	Urum (urum1249, Turkic, Eurasia: Skopeteas et al., 2024)	18797
Movima (movi1243, Isolate, South America: Haude, 2024)	10243	Vera'a (vera1241, Austronesian, Papunesia: Schnell, 2024)	17785
Dalabon (ngal1292, Gunwinyguan, Australia: Ponsonnet, 2024)	4046	Warlpiri (warl1254, Pama-Nyungan, Australia: O'Shannessy, 2024b)	7129

Table 1: The 44 languages in the DoReCo dataset. 'n tokens' is the number of target language word tokens.

target language lemmas may mismatch across seed words: English *split* might yield Spanish *ˆsepar* as a target language lemma. Without further processing, this would lead to the model's failure to recognize that *separar* translates into the reference language words *separate* and *split*.

To resolve this issue, I implement a simple heuristic to merge target language lemmas given the same or different seed words. In all cases, the basic criterion is that two target language lemmas l_i and l_j are merged iff they have a longest-common substring (1) whose length is ≥ 3 characters, and (2) that is at least half as long (in characters) as the shortest string of the two lemmas l_i and l_j. When merging *across* seed words (like *ˆsepar* given *split* and *ˆsepara* given *separate* in the example above), we further require that the whole word forms (e.g., *separamos, separaba*) that the two lemmas cover overlap, as a further way to ensure that they indeed are the same lemma. Concretely, we retrieve the set of unique whole word forms covered by l_i, i.e. all unique strings from **tokens**(l_i), and call it W_i. We do the same for l_j and call it W_j. Next, we define the two lemmas to have sufficient overlap in the word forms they cover if $|W_i \cap W_j| \geq \max(|W_i|, |W_j|) \times 0.5$, or: the intersection of their word forms is at least half the size of the largest of the two sets. All lemma pairs are considered, and an undirected graph is induced with edges between all pairs of mergeable lemmas, after which all lemmas in each connected component are merged.

2.3 Induction step

With the inferred mapping between seed words in the reference language t and merged lemmas in the target language t, we can now train a classifier to induce the merged lemma given a seed word token in the bitext between r and t. In particular, the classifier learns a mapping between contextualized vector representations $\overrightarrow{w_r}$ of each token w_r, and the merged lemmas L_t, as obtained through the previous steps.

This, then, allows for the inference of what a target language t would have used in the case of a token of some other target language t'. For every token $w_r \in B^r_{t'}$, that is: in the bitext between r and t', the contextualized vector $\overrightarrow{w_r}$ is retrieved, and **classify**$(\overrightarrow{w_r}, t)$ predicts the lemma in t for the translation of a token in t'. As such, we now know that t uses **classify**$(\overrightarrow{w_r}, t)$ for w_r, and t' uses **classify**$(\overrightarrow{w_r}, t')$ for the same token, thus making the reference language token a comparable category. Doing so for all $t \in T$ yields one row in a comparison table as obtained from a massively parallel corpus, except that most lexical labels are now inferred instead of observed. Doing so for all word tokens w_r in any bitext allows us to create the full table. I will explore the insights that can be derived from such a table in §5, but first validate the quality of this procedure.

3 Experimental set-up and materials

This paper uses the DoReCo corpus (Seifart et al., 2024), a collection of data gathered by documentary linguists for a typologically diverse sample of languages. The individual language resources form free-standing contributions that should be individually cited as part of the usage agreement. Table 1 presents the 44 languages used, along with metadata about affiliation and location and the number of (translated) words in each language.

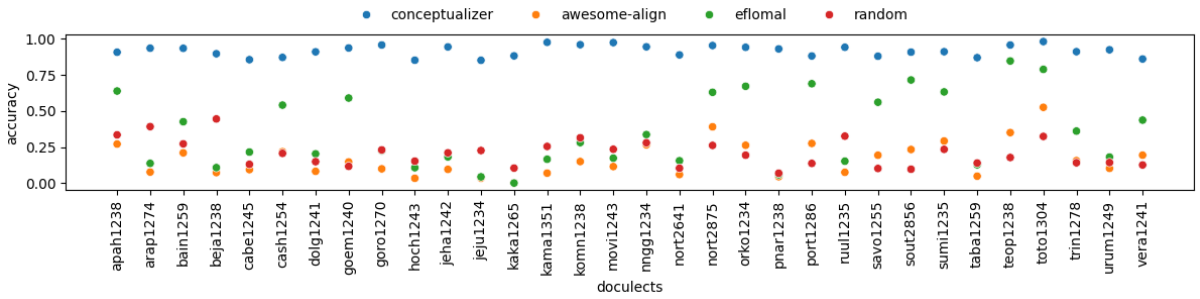


Figure 2: Mean extraction accuracy (blue) vs. random baseline (orange).

w	melo bo lo ghavilighue.
m	melo bo lo ghavi -li -ghu =e
g	tuna go 3SG.M paddle -3SG.M.O -NMLZ =EMPH
f	"he went and fished bonito with it."
l	go fish bonito

Table 2: Interlinear Gloss; Savosavo (Wegener, 2024)

The structure of the components of the DoReCo corpora is given in Table 2, for the language Savosavo. All languages have the **[w]**ord and **[f]**ree translation layer, and a select subset of languages is interlinearly glossed with a **[m]**orphological segmentation layer and a **[g]**loss layer. Subsequently, the lexical **[l]**emma layer was derived from the **f** layer, by selecting all lemmatized words from the **f** layer whose PoS was one of Noun, Adjective, or Verb, using `spacy` for both lemmatization and PoS tagging (Honnibal and Montani, 2017).

Finally, in the induction step, BERT (Devlin et al., 2019) was used, using the `bert-base-cased` model of the `transformers` library.

4 Validation experiments

This section validates the quality of the model. As the extraction of high-quality translation equivalence relations between tokens in the target and reference language is paramount for the validity of subsequent steps, I first evaluate the Liu et al. (2023) model, which provides us with such translation equivalences, in two ways: by assessing if reference language items are aligned with the correct target language tokens (§4.1), and by assessing if the extracted 'lemmas' accurately lemmatize the target language (§4.2). Next, I consider the accuracy of the lexification induction step (§4.3).

4.1 Quality of lemma extractions

To evaluate whether the correct target language tokens are aligned with the reference language word tokens, I use the glosses, available for $32/44$ DoReCo corpora. Given that the target language tokens are associated with a morphological segmentation and a corresponding gloss in English (cf. Table 2 for an example), we can assess whether the target language token aligned with a seed language item contains the seed language item as part of its gloss. For the example in Table 2, the lemma *go* (on the **[l]**exical lemma line) might be aligned with Savosavo *bo*, which is indeed glossed as 'go' (cf. the **[g]**loss line). Only reference language words that are present in at least one gloss in the target language are considered. For instance, the verb *fish* might be aligned with *ghavilighue*, but this word does not have 'fish' in one of its glosses, but rather 'paddle'. Since no other word in the target language has 'fish' in one of its glosses, the item is not counted as correct or incorrect.

We compare the scores of the Conceptualizer mode against a weak baseline of picking a word from the target language sentence at random, and a stronger baseline of a simple extraction procedure in which the alignments over word alignments obtained through either Awesome Align (Dou and Neubig, 2021) or Eflomal (Östling and Tiedemann, 2016) combined with the 'grow-diag-final-and' heuristic were used (for both models, default settings were used). The procedure furthermore involved resolving cases where one reference language word token was mapped onto multiple target language tokens, as the evaluation procedure requires a single target language form. For such cases, only the target language token that was most frequently aligned with the reference language word type across the whole bitext was kept.

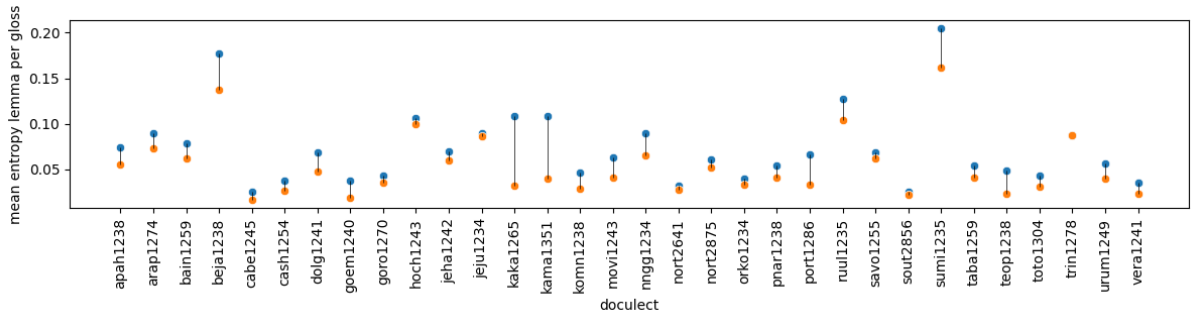


Figure 3: Entropy of lemmas given glosses. Blue: **lemma-H** without merging; orange: with merging.

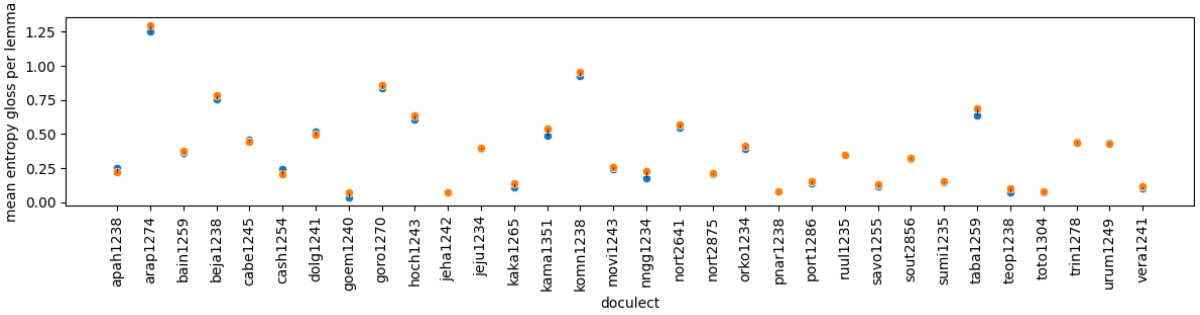


Figure 4: Entropy of glosses given lemmas. Blue: **gloss-H** without merging; orange: with merging.

Figure 2 reports the macro-averaged scores, i.e., averaged over all tokens of seed language items per language. For the Conceptualized model, the median language gets 91.1% of alignments correct, against median accuracy scores of the three baseline models of 19.7% (random), 15.5% (Awesome Align), and 25.4% (Eflomal). The variation between languages is relatively small, with an interquartile range of 88.2%-94.4%, a worst case of 81.0% (Jejuan) and a best case of 97.2% (Toto). Notably, these results substantially support the superiority of the Liu et al. (2023) Conceptualizer model over alignment-based procedures in low-resource scenarios such as the one studied here.

4.2 Effect of lemma merging

While the previous analysis supports the accuracy of the alignments between seed words and target language tokens, it does not yet validate whether the extracted lemmas, to be used in the subsequent induction step, are accurate. It may be that all target language tokens are correctly aligned, but this is done through several lemmas that all correspond to one 'true' lemma as given in the gloss. This would lead to an artificial inflation of the lexical boundaries in the language, which in turn reduces the quality of the inferred representations of crosslinguistic variation. The merger step discussed in §2.2 intends to pre-empt this situation.

It is difficult to assess the quality of the extracted lemmas directly, due to variation in how the glosses are assigned. Because of that, I approach the assessment indirectly, by considering the uncertainty in two conditional probability distributions: of extracted lemmas given annotated glosses, and, vice versa, of glosses given lemmas. I only consider gloss-lemma pairs found to be correctly aligned in the previous evaluation step.

For a target language t, let G_t be the set of all glosses that contain a seed word, i.e., the glosses used to determine the correctness of the alignment in the previous set, and L_t the set of induced lemmas (either as-is from the Liu et al. (2023) procedure, or after the merging step) found in cases of correct alignments. Primarily, I propose to measure the quality through the weighted average uncertainty of the probability of the lemmas given a gloss, or $P(L_t|g)$, for all glosses $g \in G_t$, as weighted by the frequency of occurrence of the gloss among correctly aligned cases, or $N(g)$. In an ideal case, for every gloss, there is just a single induced lemma that aligns to it. If multiple lemmas are found, aligning to the same gloss, the model might have inferred spurious lemmas. Formally, **lemma-H**$(t) =$

$$\sum_{g \in G_t} \Big(H(P(L_t|g)) \times N(g) \Big) \times \frac{1}{\sum_{g \in G_t} N(g)} \quad (1)$$

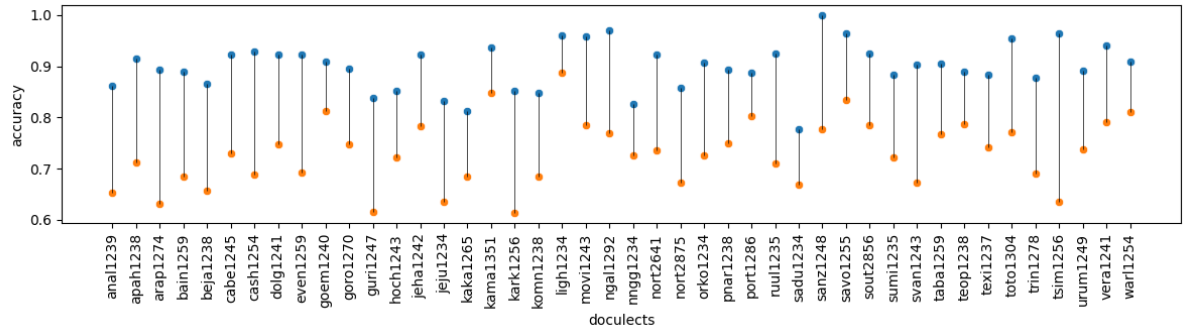


Figure 5: Accuracy on classifying the lemma for held-out data; blue: MLP-100 classifier, orange: most-frequent lemma per seed word baseline.

The inverse relation, of glosses given lemmas, similarly has an expectation of one-to-one mappings: given a lemma, we expect it to align with one unique gloss only. For this relation, however, a qualification applies: for many of the languages in the corpus, the individuable glosses contain more than just the lemma, and as such individual lemmas frequently align with multiple unique glosses, with substantial variation between the languages owing to the different approaches to writing the glosses that the documentary linguists applied. Nonetheless, the gloss entropy given the lemmas is a useful measure when assessing the effect of the merging step: if applying the merging step leads to erroneous mergers, i.e., cases where two induced lemmas are merge that should not be merged, the uncertainty over the glosses given the lemmas should go up, as the original lemmas that were erroneously merged can be expected to have rather different sets of glosses. As such, it can be expected that if the merging is accurate, the entropy over the glosses given the lemmas should *not* go up relative to the application of the model *without* the merging step. Formally, the **gloss-H** measure is defined as:

$$\sum_{l \in L_t} \Big(H(P(G_t|l)) \times N(l) \Big) \times \frac{1}{\sum_{l \in L_t} N(l)} \quad (2)$$

Figure 3 shows that across languages the **lemma-H** goes down with the addition of the merging step for each individual language, with some positive outliers being Kakabe and Kamas, where most of the uncertainty over the glosses is removed by adding the merging step (**lemma-H** values going from 0.109 to 0.032 for the former and 0.109 to 0.040 for the latter). On average, the **lemma-H** was found to decrease from 0.072 when the merging step is not applied, to 0.053 when it is applied.

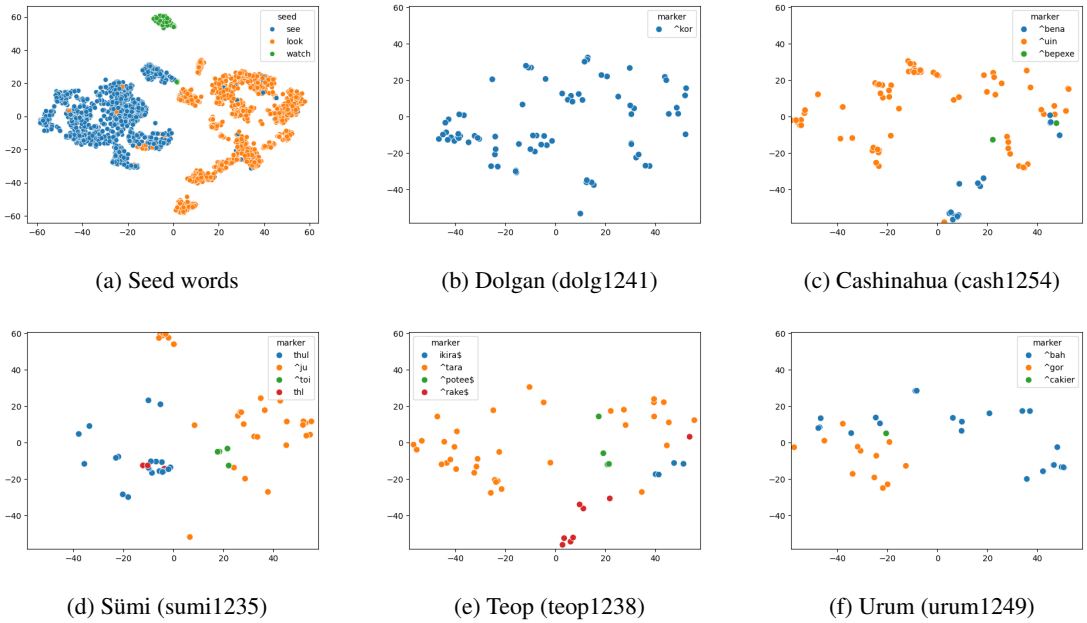
model	accuracy	ERR
baseline	0.739	-
KNN-3	0.862	0.491
SVC	0.890	0.594
MLP	0.898	0.624
MLP-100	0.900	0.631

Table 3: Induced lexification results across all languages; ERR = error rate reduction.

Conversely, the merging step does not introduce substantial new uncertainty in the $P(G_t|l)$ distributions due to erroneous lemma mergers. Compared to the magnitude of the **gloss-H** values when no merging step is applied, the **gloss-H** values when merging *is* change relatively little, as Figure 4 illustrates on a language-by-language basis. Only in 6 cases does the **gloss-H** value go up with the addition of the merging step, compared to 19 cases where it goes down, meaning that on the whole, adding the step in fact *reduces* the uncertainty over the glosses given the lemmas.

4.3 Quality of induced lexification

The two validation experiments suggest that the inferred lemmas align reasonably well with the linguistic annotations provided in the corpus. While the goal of the induction procedure is to infer the target language lemmas given contextualized usages of target words for *other* target languages, we can assess the quality of the induction procedure by assessing the classification accuracy on a held-out sample of the *same* language. For each of the 44 languages, all seed words occurring with a frequency of 10 or more were considered, and K-fold cross-validation (here: $K = 20$) over the entire lexicon of some target language t was carried out.



(a) Seed words (b) Dolgan (dolg1241) (c) Cashinahua (cash1254)

(d) Sümi (sumi1235) (e) Teop (teop1238) (f) Urum (urum1249)

Figure 6: t-SNE plots with various colour coding. For the five languages, only the observed tokens are shown.

The accuracy of this procedure was then compared against a baseline of always predicting the most common lemma given a reference language seed word, reflecting a scenario in which a model only knows the input word in the reference language. I assessed four classifiers implemented in the `sklearn` library, a k-nearest neighbours classifier with $k = 3$ (KNN-3), an Support Vector Classifier with the default settings (SVC), and two Multi-Layer Perceptron, one with no hidden layers (MLP), and the other with one hidden layer of 100 units and ReLU activation (MLP-100).

Table 3 presents the results. Seed words tend to be associated with few lemmas, one of which is typically very dominant (cf. the low entropy of the lemmas given the glosses in Figure 3, which supports this observation). As such predicting the modal lemma given a reference language seed word forms a competitive baseline. All classifiers, however, provide substantial improvement over the baseline, reducing the error by between 49% (KNN-3) and 63% (MLP-100). Figure 5 shows the results per language for the best-performing MLP-100 model, showing that the classifier surpasses the baseline and generally performs well for all languages.

5 Application

The previous section demonstrated that the model extracts generally valid target language representations (lemmas) and is reasonably well able to classify these lemmas on the basis of contextual vector representations of the seed language. The goal of this approach, however, is to provide a method for typologists to obtain massively comparable data in the absence of a massively parallel corpus. This section demonstrates how known insights can be replicated, and how novel insights can be obtained with the method.

To explore the comparability afforded by the model, here, we briefly explore the domain of visual perception verbs, translation equivalents of English *see, look,* and *watch*. A main lexical distinction between Experiencer and Activity verbs (English *see* vs. *look*) – with the former involving a more passive ('experiencing') role for the perceiver, and the latter a more active one has been postulated (Viberg, 1983), but challenged on the basis of parallel corpus data by Wälchli (2016). Using manually extracted instances from comparable corpora, San Roque et al. (2018) consider the non-literal extensions of perception verbs, noting that discourse markers (e.g., *look!* to draw attention or introduce something unexpected) are common extensions.

To explore the distribution of visual perception verbs in the DoReCo corpus, we can train the best classifier from §4.3 (MLP-100) for each language that has $N \geq 30$ instances of the three most common English visual perception verbs (*see, look,* and *watch*) in their free translations. Next, we apply

this classifier to all instances of *see, look*, and *watch* for all other languages, leading to a 3001 (instances of visual perception verbs across all 29 languages with sufficient data) by 194 (unique lemmas across the 37 languages with sufficient data) table, with the probability assigned by each MLP-100 model to the lemmas as the cell values. To visualize this table, we can apply t-SNE (Van der Maaten and Hinton, 2008) to reduce the table to two dimensions.

Figure 6 shows the t-SNE representation, with six distinct colour-codings. The top-left subfigure (6a) shows the distribution of the three English seed words, which form coherent groups of visual clusters, but with each term nonetheless covering multiple clusters. Some languages, such as Dolgan (6b) do not make any lexical distinctions in this domain – a situation predicted by Viberg (1983)), while others, such as Sümi (6d) split Activity and Experiencer meanings more or less along the lines of English. Two languages carve out a cluster near the bottom of the 2D-space – Cashinahua *bena* and Teop *rake* – these are all instances of *look for*, meaning 'search', which many languages group with the other 'look' meanings, but these two languages distinguish lexically. Finally, Urum (6f) presents an interesting case of two main terms, but with a split that differs from English or Sümi. Here, we see that *bah* covers a region containing English *look* and some of *see*, whereas *gor* covers only part of the *see* tokens. The *see* tokens covered by *bah* involve cases of modal see, like *can see*, *will see*, in several cases in the meaning 'find out', like "I will see where to go, possibly to the city". As such, Urum supports the argument of Wälchli (2016) that the Activity-Experiencer split is (a) more of a continuum, and (b) governed by properties beyond the general semantic role of the perceiver.

What the plots in Figure 6 further illustrate, is that languages differ in how often they use visual perception meanings. Urum uses visual perception verbs only 21 times per $10,000$ tokens, whereas Cashinahua shows five times that frequency at 102 tokens per $10,000$. Such usage variation is known to be meaningful in the explanation of lexification patterns, following the argument that a language's greater need to communicate about a specific concept correlates with finer-grained lexical distinctions (cf. Kemp et al., 2018). Original corpus data and methods for making such data comparable can thus be used to estimate such 'need probabilities'

6 Conclusion

This paper introduced a novel method for making original text corpora that are translated into the same reference language comparable, thus allowing for token-level typological study. The independent steps of the method were found to generally provide high-quality representations in three validation experiments, and the case study presented the potential of the method for studying lexical semantic variation across languages.

While generally successful in extracting translation equivalents and inducing lexical categorization models, room for improvement remains. While the Liu et al. (2023) approach benefits from its ability to consider substrings below the word level, it is hampered by not considering how other target language substrings translate to the seed item, something word alignment procedures from IBM-1 (Brown et al., 1993) onward do consider.

It should be stressed here that using original text does not make the method bias-free, in terms of a translationese bias from the shared reference language. Using the free translations means all lexical choice models are filtered through contextual vector representation of English. In the specific case of the data used here, this English is moreover written as a guide for the linguistically informed reader to make sense of the target language sentence; it may, by design given the genre of "free translations in language documentation", show translation effects from the target language onto the English. Calibrating the extent of this effect would require further testing the model on other comparable corpora.

Applications beyond the ones the method was designed for could be explored. Related work that considers crosslinguistic variation at a word type level and using secondary resources, like Thompson et al. (2020) and Khishigsuren et al. (2025), could be compared against the token-level mappings between a shared reference language and multiple target languages. Corpora that contain both original and translated text in comparable genres may furthermore be of use to pinpoint the precise effects of translationese in how lexical boundaries are drawn, and as such be of use for practical purposes in education and translation studies. Finally, we are reminded that languages vary on a discourse-pragmatic level, and that multilingual NLP ought to consider such variation, for instance when working with Large Language Models and Machine Translation systems pretrained on translated text.

References

Maria Bardají i Farré. 2024. Totoli DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Natalia Bogomolova, Dmitry Ganenkov, and Nils Norman Schiborr. 2024. Tabasaran DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Niclas Burenhult. 2024. Jahai DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Alexander Yao Cobbinah. 2024. Baïnounk Gubëeher DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Andrew Cowell. 2024. Arapaho DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pedro Henrique Domingues, Claudio Santos Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. Quantifying the ethical dilemma of using culturally toxic training data in AI tools for indigenous languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 283–293, Torino, Italia. ELRA and ICCL.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Chris Lasse Däbritz, Nina Kudryakova, Eugénie Stapert, and Alexandre Arkhipov. 2024. Dolgan DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Christian Döhler. 2024. Komnzo DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Renata Enghels, Bart Defrancq, and Marlies Jansegers. 2020. *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges*. Walter de Gruyter GmbH & Co KG.

Diana Forker and Nils Norman Schiborr. 2024. Sanzhi Dargwa DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Michael Franjieh. 2024. Fanbyak DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Alexandro Garcia-Laguia. 2024. Northern Alta DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Jost Gippert. 2024. Svan DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Richard Griscom. 2024. Asimjeeg Datooga DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Valentin Gusev, Tiina Klooster, Beáta Wagner-Nagy, and Alexandre Arkhipov. 2024. Kamas DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Tom Güldemann, Martina Ernszt, Sven Siegmund, and Alena Witzlack-Makarevich. 2024. Nng DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Geoff Haig, Maria Vollmer, and Hanna Thiele. 2024. Northern Kurdish (Kurmanji) DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Iren Hartmann. 2024. Hoocąk DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Andrew Harvey. 2024. Gorwaa DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Martin Haspelmath. 2018. How comparative concepts and descriptive linguistic categories are different. In Daniël Olmen, Tanja Mortelmans, and Frank Brisard, editors, *Aspects of linguistic variation*, pages 83–114. De Gruyter Mouton.

Katharina Haude. 2024. Movima DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Birgit Hellwig. 2024. Goemai DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Birgit Hellwig, Gertrud Schneider-Blum, and Khaleel Bakheet Khaleel Ismail. 2024. Tabaq (Karko) DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Stig Johansson and Knut Hofland. 1994. Towards an english-norwegian parallel corpus. In G. Tottie Fries and P. Schneider, editors, *Creating and using English language corpora*, pages 25–37. Rodopi, Amsterdam.

Olga Kazakevich and Elena Klyachko. 2024. Evenki DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1).

Temuulen Khishigsuren, Terry Regier, Ekaterina Vylomova, and Charles Kemp. 2025. A computational analysis of lexical elaboration across languages. *Proceedings of the National Academy of Sciences*, 122(15):e2417304122.

Soung-U Kim. 2024. Jejuan DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Manfred Krifka. 2024. Daakie DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Natalia Levshina. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica*, 50(2):507–542.

Natalia Levshina. 2017. Online film subtitles as a corpus: An n-gram approach. *Corpora*, 12(3):311–338.

Natalia Levshina. 2021. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, pages 129–160.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schütze. 2023. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12969–13000.

Anthony McEnery and Zhonghua Xiao. 2007. Parallel and comparable corpora: What are they up to. *Incorporating corpora: translation and the linguist*, pages 18–31.

Felicity Meakins. 2024. Gurindji DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Ulrike Mosel. 2024. Teop DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Carmel O'Shannessy. 2024a. Light Warlpiri DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Carmel O'Shannessy. 2024b. Warlpiri DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Pavel Ozerov. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Claudio S. Pinhanez, Paulo Cavalin, Marisa Vasconcelos, and Julio Nogima. 2023. Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6174–6182. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Maïa Ponsonnet. 2024. Dalabon DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Juan Diego Quesada, Stavros Skopeteas, Carolina Pasamonik, Carolin Brokmann, and Florian Fischer. 2024. Cabécar DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Sabine Reiter. 2024. Cashinahua DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Sonja Riesberg. 2024. Yali (Apahapsili) DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Hiram Ring. 2024. Pnar DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Françoise Rose. 2024. Mojeño Trinitario DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Lila San Roque, Kobin H Kendrick, Elisabeth Norcliffe, and Asifa Majid. 2018. Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics*, 29(3):371–406.

Stefan Schnell. 2024. Vera'a doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Frank Seifart, Ludger Paschen, and Matthew Stave. 2024. *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Stavros Skopeteas, Violeta Moisidi, Nutsa Tsetereli, Johanna Lorenz, and Stefanie Schröter. 2024. Urum DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Deborah Tannen. 1980. A comparative analysis of oral narrative strategies: Athenian Greek and American English. In Wallace L. Chafe, editor, *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*, pages 51–87. Ablex Publishing Company Norwood^ eN. J NJ.

Amos Teo. 2024. Sümi DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Marina Terkourafi. 2011. The pragmatic variable: Toward a procedural interpretation. *Language in Society*, 40(3):343–372.

Nick Thieberger. 2024. Nafsan (South Efate) DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Martine Vanhove. 2024. Beja DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Annemarie Verkerk. 2014. Where Alice fell into: Motion events from a parallel corpus. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating*

dialectology, typology, and register analysis: Linguistic variation in text and speech, pages 324–354.

Åke Viberg. 1983. The verbs of perception: A typological study.

Alexandra Vydrina. 2024. Kakabe DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Bernhard Wälchli. 2016. Non-specific, specific and obscured perception verbs in Baltic languages. *Baltic Linguistics*, 7:53–135.

Claudia Wegener. 2024. Savosavo DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Søren Wichmann. 2024. Texistepec Popoluca DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Alena Witzlack-Makarevich, Saudah Namyalo, Anatol Kiriggwajjo, and Zarina Molochieva. 2024. Ruuli DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Xianming Xu and Bibo Bai. 2024. Sadu DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Are Translated Texts Useful for Gradient Word Order Extraction?

Amanda Kann
Stockholm University
Stockholm, Sweden
amanda.kann@su.se

Abstract

Gradient, token-level measures of word order preferences within a language are useful both for cross-linguistic comparison in linguistic typology and for multilingual NLP applications. However, such measures might not be representative of general language use when extracted from translated corpora, due to noise introduced by structural effects of translation. We attempt to quantify this uncertainty in a case study of subject/verb order statistics extracted from a parallel corpus of parliamentary speeches in 21 European languages. We find that word order proportions in translated texts generally resemble those extracted from non-translated texts, but tend to skew somewhat toward the dominant word order of the target language. We also investigate the potential presence of underlying source language-specific effects, but find that they do not sufficiently explain the variation across translations.

1 Introduction

When investigating cross-lingual transfer in multilingual language models, NLP researchers often rely heavily on data from typological databases such as WALS (Dryer and Haspelmath, 2013) for quantitative measures of language distance.[1] These databases typically reduce cross-linguistic variation to a set of categorical binary distinctions, obscuring the intra-linguistic variation present in many features (Wälchli, 2009), including word order.

This type of gradient variation is better captured by continuous token-level measures, such as statistical distributions of specific constructions (e.g. individual word order types) observed in annotated corpora (Levshina et al., 2023; Baylor et al., 2024). Corpus-based measures also allow for greater transparency and reproducibility than manual categorical judgments, and enable cross-linguistic comparisons at the potential scale of thousands of languages with maintained methodological consistency (see e.g. Östling and Kurfalı, 2023).

However, care must be taken to ensure that the selected texts are both sufficiently representative of their respective languages and comparable across languages, in order to control for variation resulting from differences between text types. Using massively parallel texts ensures that text type and pragmatic context will be identical across all analyzed languages, reducing the risk of misleading cross-linguistic comparisons (Ebert et al., 2024).

Parallel texts are also inherently translational, however, and could thus diverge structurally from original (non-translated) texts because of artefacts introduced in the translation process. For instance, translated texts commonly contain less lexical and grammatical variation than original texts in the same language (*regularization*). Structural properties of the source language may also be retained in translation, even when they are marked in the target language (*source language interference*). Thorough descriptions of features theorized to be cross-linguistically typical of translated text can be found in translation studies literature (e.g. Baker, 1993).

Translational artefacts can be strong enough to train reliable classifiers for automatic detection of translated texts (Volansky et al., 2015), and to accurately determine the relative genealogical distance between different source-target language pairs based only on cues in translations (Rabinovich et al., 2017). The cited studies rely heavily on syntactic features (most commonly part-of-speech *n*-grams), suggesting that translational artefacts could have a direct impact on word order proportions – however, word order (particularly of subject, verb and object) is not necessarily well captured by part-of-speech sequences, and the relationship between general and source-language specific translation effects in this domain has yet to

[1] For an overview of common approaches and typological distance measures in cross-lingual transfer research, see Philippy et al. (2023).

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 177–182
August 1, 2025 ©2025 Association for Computational Linguistics

be systematically studied.

We therefore conduct an analysis of translational artefacts in gradient subject/verb order extraction from a parallel corpus of transcribed speeches with high-quality human translations in 21 languages. Our aim is to investigate:

- whether gradient word order statistics extracted from translations vary significantly from those extracted from original texts, and
- whether the direction or amplitude of such differences is influenced by word order preferences in the source language.

We expect that observed variation will be stronger in the direction of the dominant word order (as a result of regularization), and that source language interference will pull the word order proportions of translations toward the proportions observed in their source texts.

2 Data

We use *CoStEP* (Graën et al., 2014), a cleaned and turn-level aligned version of the *Europarl* parallel corpus (Koehn, 2005). *Europarl* consists of transcribed speeches and human translations in 21 European languages, obtained from European Parliament proceedings between 1996 and 2011. Since both the original speeches and their translations are present in the corpus, the source language for any given translated sentence is always known – this quality is essential for disambiguating potential source language-specific effects. All 420 possible source-target language pairs occur in the corpus, with data sizes ranging between 21 885 (Estonian–Bulgarian) and 8 738 402 (English–French) tokens. The corpus contains considerably more text (both original and translated) in the 11 languages that already had official EU language status prior to the expansions in 2004 and 2007.

To enable syntactic analysis, all texts (both original and translated, across all 21 languages) have been automatically tokenized, part-of-speech tagged and dependency parsed using the monolingual *Universal Dependencies* (Nivre et al., 2020) models availale through `Stanza` (Qi et al., 2020). While the parsing accuracy of these models varies somewhat across languages, noise from automatic annotation appears to have a minimal impact on word order proportions extracted from larger corpora (Levshina et al., 2023) – in addition, cross-linguistic performance differences do not directly affect comparisons between translations into the same language (regardless of source language).

3 Word order extraction

Subject/verb order can be defined and delimited in several ways, capturing different constructions and patterns of variation. We use a combination of part-of-speech and dependency tags on a given token and its direct head, operationalizing the relative order of nominal subject and verb as [NOUN|PROPN] $\overleftarrow{\texttt{nsubj}}$ [VERB] (i.e. a nominal subject relation between a noun or proper noun and a verb). Following Ebert et al. (2024), we only consider main clauses, and in auxiliary constructions we use the position of the finite verb (which may be an auxiliary) rather than the lexical verb. We include both transitive and intransitive verbs, and both declaratives and interrogatives; however, we distinguish these categories in extraction so that they can be analyzed separately.

We split the corpus by target language and compute the relative frequencies of both possible word orders (subject-verb and verb-subject) separately per source language.[2] The resulting word order proportions for each source-target pair are then compared to the reference proportion extracted from original texts in the target language.

4 General translation effects

Figure 1 displays the distributions of verb-subject (VS) order proportions per language pair, grouped by target language and sorted by VS proportion in original texts in the target language. All languages in the corpus prefer subject-verb (SV) order[3], to varying degrees. The highest VS proportions are found in German (`de`), Estonian (`et`), Swedish (`sv`) and Dutch (`nl`); this is expected, as their dominant word order in main clauses is typically analyzed as *verb-second* (or, for spoken Estonian, *verb-third*) rather than SV (Vihman and Walkden, 2021).

Overall, the proportions observed in translated texts are similar to original texts – the mean difference across language pairs is -0.017. However, there is also variation between translated texts with different source languages. Even for French (`fr`), which has the lowest dispersion across translations

[2] Following Levshina et al. (2023), we set a minimum total frequency threshold of 500 occurrences of the construction of interest – 412 of 420 language pairs in the corpus meet this threshold for nominal subject/verb constructions.

[3] This preference is expected for all languages in the *Europarl* sample; see section 6 for further discussion.

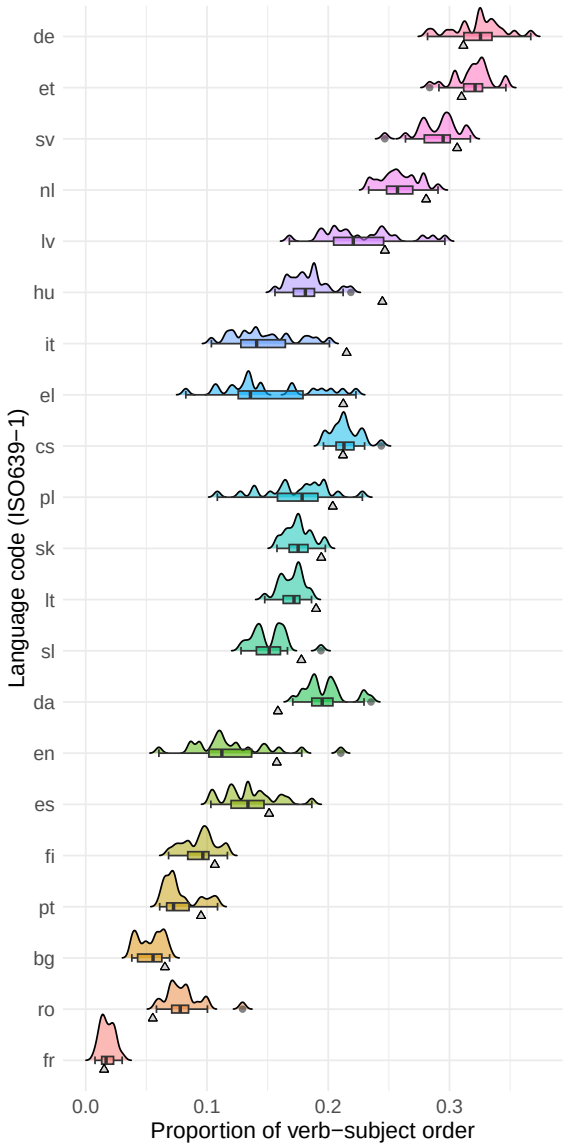


Figure 1: Distributions of VS order proportions in translated *Europarl* texts with different source languages, grouped by target language. The box plots display the median, interquartile range and whiskers extending to 1.5 IQR, with outliers plotted as individual points. The gray triangles indicate the VS order proportion in original texts in the respective target language.

($n_t = 20; \sigma_t = 0.0057; IQR_t = [0.013, 0.023]$), grouping the original French data by year of production (as a non-translational reference variable) results in a distribution with slightly lower dispersion ($n_{year} = 17; \sigma_{year} = 0.0027; IQR_{year} = [0.013, 0.017]$). Similar results are found for German (de), suggesting the presence of some unexplained variation specific to translated texts.

It should be noted that this variation is of a similar scale to the differences resulting from operationalizing the word order of interest differently; for instance, including only intransitive sentences results in higher dispersion for both the translations ($\sigma_{t_{Intr}} = 0.0077$) and the reference population ($\sigma_{year_{Intr}} = 0.0038$).

For 15 of 21 languages in the sample, the VS proportion in original texts is higher than both the median and upper quartile of VS proportions in the population of translations into that language; several original texts (e.g. Italian (it) and Hungarian (hu) would be outliers in their respective populations. The overall population of differences in VS proportion between translations and original texts (across all target languages) is approximately normally distributed, with a slight negative skew ($\tilde{x} = -0.015, IQR = -0.034, 0.004$). This tendency toward SV order in translations aligns with our hypothesis, and may be a reflection of the regularization effects described in section 1.

5 Source language-specific effects

To examine the potential effects of source language interference, VS order proportions from the set of translated turns in a given source-target language pair are also compared to the proportions extracted from the same turn set in the source language. Figure 2 plots this relationship for all source languages, into three target languages with different mean VS order proportions and dispersions across translations. We find no signifi-

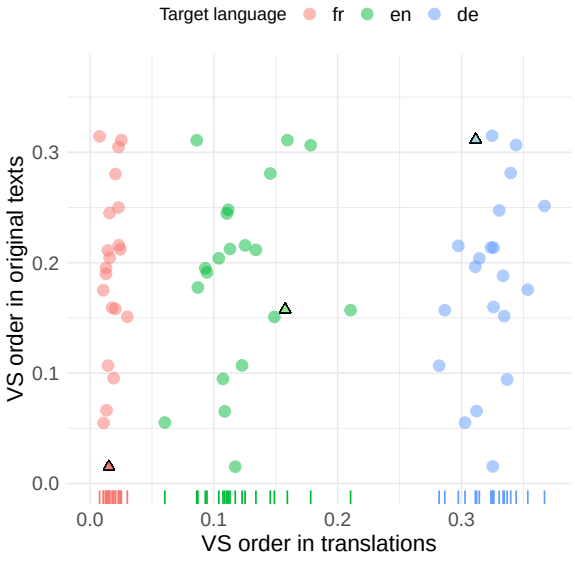


Figure 2: VS order proportions in translations (into French (fr), English (en) and German (de)) and in original texts for each language. The triangles indicate the proportions in original texts in the target languages.

cant correlations between source and target proportions for any individual target language, as would have been expected from the hypothesized source language interference model. Across the entire population, we find a weak positive correlation between source proportion and translation effect (the difference between proportions in translations and original texts in the target language), but it explains very little of the variation in translation effects ($\beta = 0.08; CI_{95\%,\beta} = [0.04, 0.12]$; $R^2 = 0.04; CI_{95\%,R^2} = [0.01, 0.08]$).

A potentially confounding source language effect is the proportion of different clause types in the source texts (assuming that they are carried over to the target language), since word order preferences for different clause types vary across languages. For instance, SV order in transitive clauses is stricter than in intransitive clauses in some European languages, such as Spanish and Latvian (Dryer, 2013) – in these languages, the proportion of transitive clauses should have a greater impact on the extracted word order proportion than in languages where the two clause types pattern similarly. The *Europarl* data supports this claim: there is a positive correlation between the proportion of intransitive clauses and VS order proportion in both Spanish and Latvian. However, we unexpectedly find no correlation between intransitive clause proportions in source texts and translations, either for these languages or across the entire sample. While the turn-level alignment of *CoStEP* is too coarse to meaningfully investigate this further, individual clause-level comparison in a word-aligned parallel corpus could verify to what extent properties of source language clauses which may influence word order proportions are preserved in translation.

6 Conclusions

In this study, we analyzed the general and source language-specific effects of translation on verb/subject order statistics extracted from *Europarl*. We observed a general tendency toward rigid SV order in translations compared to original texts, in line with the broader *regularization* effect discussed in translation studies literature. Unexpectedly, we found that word order proportions in the source texts do not sufficiently explain this tendency, at least when averaged at corpus level. This suggests that controlling for source language factors will not reliably reduce uncertainty when using translated texts to approximate word order distributions in original texts.

Crucially, the issue of translational artefacts should not disqualify good-quality translations from use in the extraction of gradient word order typology, assuming that the uncertainty in the extracted proportions is properly taken into account in interpretation – as is good practice for any parameter by which syntactic properties of a text may vary. As with text genres, including multiple different source languages in a corpus of translations may reduce the risk of unrepresentativity. A well-motivated theoretical definition (and operationalization) of the word order feature of interest is also necessary in order to make valid cross-linguistic comparisons based on extracted word order proportions. With these aspects in mind, even an uncertain estimate of gradient word order proportions will encode considerably more fine-grained and useful comparative information than the customary binary word order classifications.

It is important to note the restricted scope of this case study. We only investigate one word order feature, which is particularly prone to pragmatically motivated variation in many languages. Additionally, the language sample in *Europarl* is highly areally and genealogically skewed. Most languages in the sample are members of the Standard Average European *Sprachbund*, and are thus likely to share some cross-linguistically marked syntactic features – for instance, inverted subject/verb order in polar questions (Haspelmath, 2001). *Europarl* is also unusual in other aspects, such as text genre (formal speeches, with higher average sentence and utterance length than spontaneous informal speech) and the purpose of translation (accurate representation of the original speeches, likely prioritizing clear language). These properties should be kept in mind when applying our findings to other contexts.

We hope that this study can serve as a framework for further cross-lingual investigations of the effects of translation on word order. In addition to analyzing more word order features, future work could cover a larger and more diverse language sample by making use of machine translations, which are an interesting object of analysis in their own right. Machine translations appear to produce different translational artefacts to human translations (Bizzoni et al., 2020), and – not least because of the prevalence of machine translated text in large text datasets – a comparison between word order extractions from human and machine translations would be very useful.

Limitations

In addition to the areal and genealogical bias discussed in section 6, the sample in *Europarl* consists entirely of high-resource languages. Accurate pre-trained parsing models are only available for a fraction of the world's languages (`Stanza` provides UD models for fewer than 100 languages), and high quality training data for PoS tagging and dependency parsing is similarly scarce.

Our word order extraction method is simple, and the per-text average measure obscures the various underlying causes of potential word order variation. Subject/verb order preferences can vary structurally across clause types or nominal categories, or pragmatically for information structure or discourse reasons – this method can only disambiguate between the structural variation sources which are accounted for in the chosen word order operationalization.

Finally, the analysis of source language-specific effects is complicated by the potential presence of indirect translations (where an intermediate language is used in the translation process). Ustaszewski (2021) reports that translations in *Europarl* produced after the official EU language expansion in 2004 more likely use an intermediate language (most commonly English), while earlier translations are more likely direct. The general impact of an intermediate language on the presence of source language artefacts in translations is unclear and warrants further investigation.

Acknowledgments

This work was made possible by individual PhD student funding from the Department of Linguistics at Stockholm University. We thank Bernhard Wälchli and Robert Östling for their valuable comments on prior versions of this paper.

Supplementary materials

The code used to produce the results and figures presented in this paper is available at `https://github.com/amandakann/sigtyp2025`, under the GPL-3.0 license.

References

Mona Baker. 1993. Corpus Linguistics and Translation Studies — Implications and Applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology*, pages 9–24. John Benjamins Publishing Company, Amsterdam.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual Gradient Word-Order Typology from Universal Dependencies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–49, St. Julian's, Malta. Association for Computational Linguistics.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Matthew S. Dryer. 2013. Order of Subject and Verb (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo. Type: Data set.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Christian Ebert, Balthasar Bickel, and Paul Widmer. 2024. Areal and phylogenetic dimensions of word order variation in Indo-European languages. *Linguistics*, 62(5):1085–1116.

Johannes Graën, Dolores Batinić, and Martin Volk. 2014. Cleaning the Europarl Corpus for Linguistic Applications. In *Konvens 2014*, Hildesheim. Stiftung Universität Hildesheim.

Martin Haspelmath. 2001. The European linguistic area: Standard Average European. In Martin Haspelmath, Ekerhard König, Wulf Oesterreicher, and Wolfgang Raible, editors, *Language Typology and Language Universals*, number 20/2 in Handbücher zur Sprach- und Kommunikationswissenschaft [HSK], pages 1492–1510. De Gruyter Mouton, Berlin, New York.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources*

and Evaluation Conference (LREC'20), pages 4034–4043, Marseille, France. European Language Resources Association (ELRA).

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in Translation: Reconstructing Phylogenetic Language Trees from Translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Michael Ustaszewski. 2021. Towards a machine learning approach to the analysis of indirect translation. *Translation Studies*, 14(3):313–331.

Virve-Anneli Vihman and George Walkden. 2021. Verb-second in spoken and written Estonian. *Glossa: a journal of general linguistics*, 6(1):15.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Bernhard Wälchli. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology*, 13(1):77–94.

Robert Östling and Murathan Kurfalı. 2023. Language Embeddings Sometimes Contain Typological Generalizations. *Computational Linguistics*, 49(4):1–49.

Author Index